

Project no. 269977

**APARSEN**  
**Alliance for Permanent Access to the Records of Science  
Network**

**Instrument:** Network of Excellence

**Thematic Priority:** ICT 6-4.1 – Digital Libraries and Digital Preservation

**D26.1 REPORT AND STRATEGY ON  
ANNOTATION, REPUTATION AND  
DATA QUALITY**



---

Document identifier:	<b>APARSEN-REP-D26_1-01-1_0</b>
Due Date:	2012-02-28
Submission Date:	2012-10-31
Work package:	WP26
Partners:	AWI, CERN, ESA, FORTH, GLOBIT, KNAW-DANS, SBA, STFC
WP Lead Partner:	Alfred Wegener Institute
Document status	Released
URN	urn:nbn:de:101-20140516191

---

### Abstract:

APARSEN is a Network of Excellence that aims to bring together an extremely diverse set of practitioner organisations and researchers in order to bring coherence, cohesion and continuity to research into barriers to the long-term accessibility and usability of data, by exploiting the project partners' diversity by building a long-lived Virtual Centre of Digital Preservation Excellence.

The broad topic of annotation, reputation and data quality is an important area in the context of the permanent access to scientific data, not least, because excellent science can build only on high or at least known quality data. Annotation, in general terms, is information added to data and it may be argued that all kinds of metadata are special types of annotation. High quality data is an essential condition of excellent science.

There are strong relations between annotation, reputation and data quality: For example, high quality should result in high reputation; high reputation should permit to assume high quality. In the context of preservation this suggests that the data must be annotated concerning quality and reputation, to facilitate appraisal (whether and what to preserve) and re-use (establish reliability). Establishing broad and deep context, detailed critique and amendment through annotation contributes further to establish trust in data, enables their review, being added to or being merged into larger scale datasets and finally, higher level data products being derived from them.

It may be argued that all kinds of metadata are just special types of annotations. This report will not much delve into the matter of metadata but concentrate on other kinds of annotation (e.g. annotation for 'comprehension and study', 'interpretation and divulgation', 'cooperation and revision'), outlined in chapter 1.1. The chapters 1.2 and 1.3 on reputation and quality deal with the status of these special topics, hinting at their relation with preservation. Chapter 2 and its corresponding sub-chapters give an overview of perspectives on and some innovative research examples on these three subjects.

Chapter 3 presents results of an internal survey to show how partners of the APARSEN network deal with annotation, reputation and data quality today.

Finally, based on a critical evaluation of the findings, a research strategy is described in general and for each of the three subjects.

**In-line with the DoW, this report does not contain new, original research findings by this project nor can it already contain a final consensus on the alignment of the partners' research agendas. It focuses on trustworthy records of research data.**

<b>Delivery Type</b>	Report
<b>Author(s)</b>	Hans Pfeiffenberger, Heinz Pampel, Angela Schäfer, Veronica Guidetti, Christoph Bruch, Yannis Tzitzikas, Stefan Pröll, Rudolf Mayer, Salvatore Mele, Patricia Herterich, Sünje Dallmeier-Tiessen, Marjan Grootfeld, René van Horik
<b>Approval</b>	Simon Lambert
<b>Summary</b>	
<b>Keyword List</b>	Data quality, Data annotation
<b>Availability</b>	<input checked="" type="checkbox"/> Public

Document Status Sheet

Issue	Date	Comment	Author
0_0	14 Jan 2012	First draft	Hans Pfeiffenberger, Heinz Pampel, Angela Schäfer, AFPUM
0_1	12 Apr 2012	First draft + ESA Input	Veronica Guidetti, ESA
0_2	14 May 2012	change of structure	Christoph Bruch, Hans Pfeiffenberger, AFPUM
0_3	06 Jun 2012	additions	Yannis Tzitzikas, FORTH-ICS Stefan Proell, Rudolf Mayer, SBA Salvatore Mele, Patricia Herterich, Sünje Dallmeier-Tiessen, CERN
0_4	22 June 2012	additions; change of structure	Hans Pfeiffenberger, AFPUM Marjan Grootfeld, DANS
0_5	25 July 2012	insertion of additions from Hans' version from 6 July 2012	Christoph Bruch, AFPUM
0_6	14 Aug 2012	confirmation of most of the editions still in change mode	Christoph Bruch, AFPUM
0_7	04 Sept 2012	Text review. Minor adjustments	René van Horik, DANS
0_8	28 Sept 2012	Minor adjustments	Christoph Bruch, AFPUM
0_9	08 Oct 2012	Some additions and adjustments, Review of “research agenda”	Hans Pfeiffenberger, AFPUM
0_10	29 Oct 2012	Internal review	Heikki Helin
1_0	31 Oct 2012	Post-review with minor additions	Yannis Tzitzikas, FORTH-ICS

**Project information**

Project acronym:	<b>APARSEN</b>
Project full title:	<b>Alliance for Permanent Access to the Records of Science Network</b>
Proposal/Contract no.:	<b>269977</b>

**Project Co-ordinator: Simon Lambert/David Giaretta**

Address:	STFC, Rutherford Appleton Laboratory Chilton, Didcot, Oxon OX11 0QX, UK
Phone:	+44 1235 446235
Fax:	+44 1235 446362
Mobile:	+44 (0) 7770326304
E-mail:	Simon.lambert@stfc.ac.uk / david.giaretta@stfc.ac.uk

## CONTENT

<b>1</b>	<b>ANNOTATION, REPUTATION AND DATA QUALITY: CONCEPT AND STATUS .....</b>	<b>7</b>
1.1	Annotation.....	7
1.1.1	User-generated annotations.....	10
1.1.2	Automatic annotation.....	11
1.1.3	Metadata in Digital Preservation .....	12
1.1.4	References and further reading .....	12
1.2	Reputation .....	13
1.2.1	Citation, persistent identifiers and research data repositories .....	13
1.2.2	Research data journals .....	14
1.2.3	Commenting and recommendation functions.....	15
1.2.4	Visibility.....	16
1.3	Data quality .....	16
1.3.1	Interweaving publications and research data .....	17
1.3.1.1	Peer review and reproducibility .....	18
1.3.1.2	Enhanced publications, executable papers and rich internet publications.....	19
1.3.1.3	Problems, challenges and chances .....	19
1.3.2	Reproducibility by using scientific workflow management systems.....	20
1.3.3	Long term preservation of scientific experiments .....	23
1.3.4	Conclusions and outlook of executable publications and scientific workflow management systems.....	24
<b>2</b>	<b>ANNOTATION, REPUTATION AND DATA QUALITY: APPROACHES BY APARSEN PARTNERS.....</b>	<b>25</b>
2.1	Annotation and annotation services .....	25
2.1.1	Example: Alfalab (DANS).....	25
2.1.2	Example: Data Preservation in High Energy Physics" (DPHEP).....	26
2.1.2.1	Activity: HEPData integration in INSPIRE.....	26
2.1.2.2	References and further reading: .....	28
2.1.3	Example: Research and development activities of FORTH-ICS.....	28
2.2	Reputation .....	32
2.2.1	Example: STD-DOI and DataCite (Helmholtz Association).....	32
2.2.2	Example: ODE project (several APARSEN partners) .....	33
2.2.3	Example: EASY (DANS) .....	34
2.3	Data quality .....	34
2.3.1	Example: MaNIDA - Coordinated provision of quality information (Helmholtz-Association) .....	35
2.3.2	Example: Earth System Science Data Journal (Helmholtz-Association).....	38
2.3.3	Example: Remote sensing domain (ESA).....	39
2.3.3.1	Context.....	39
2.3.3.2	Activity.....	40
2.3.3.3	Useful references .....	41
<b>3</b>	<b>RESULTS OF AN INTERNAL SURVEY .....</b>	<b>42</b>
3.1	Data repositories landscape .....	43
3.1.1	Data repository systems.....	44
3.1.2	Ingest and accessibility.....	47
3.2	Annotation services .....	48
3.3	Reputation .....	51
3.4	Data quality .....	53
<b>4</b>	<b>ANALYSIS AND UPCOMING RESEARCH STRATEGIES.....</b>	<b>56</b>
4.1	Critical analysis .....	56
4.2	Research strategies .....	57
4.2.1	Annotation and annotation services.....	58
4.2.2	Reputation.....	59
4.2.3	Data quality .....	59

<b>5</b>	<b>CONCLUSIONS AND RECOMMENDATIONS .....</b>	<b>60</b>
<b>6</b>	<b>LIST OF FIGURES .....</b>	<b>61</b>
<b>7</b>	<b>LIST OF TABLES .....</b>	<b>61</b>
<b>8</b>	<b>LIST OF ILLUSTRATIONS .....</b>	<b>61</b>
<b>9</b>	<b>REFERENCES .....</b>	<b>62</b>
<b>10</b>	<b>ANNEX: LIST OF QUESTIONS .....</b>	<b>64</b>

## 1 Annotation, reputation and data quality: Concept and status

The broad topic of annotation – in general terms: information added to data –, reputation and data quality is an important area in the context of the permanent access to scientific data, not least, because excellent science can build only on high or at least known quality data.

There are strong relations between annotation, reputation and data quality: For example, high quality should result in high reputation; high reputation should permit to assume high quality. In the context of preservation this suggests that there must be annotation to the data about quality and reputation, to facilitate appraisal (whether and what to preserve) and re-use (establish reliability).

Establishing broad and deep context, detailed critique and amendment through annotation contributes further to establish trust in data, enables their review, their ability to be added or merged into larger scale datasets and finally, to derive higher level data products from them.

It may be argued that all kinds of metadata are just special types of annotations. This report will not much delve into the matter of metadata but concentrate on other kinds of annotation, such as annotation for 'comprehension and study', 'interpretation and divulgation', 'cooperation' and 'revision'.

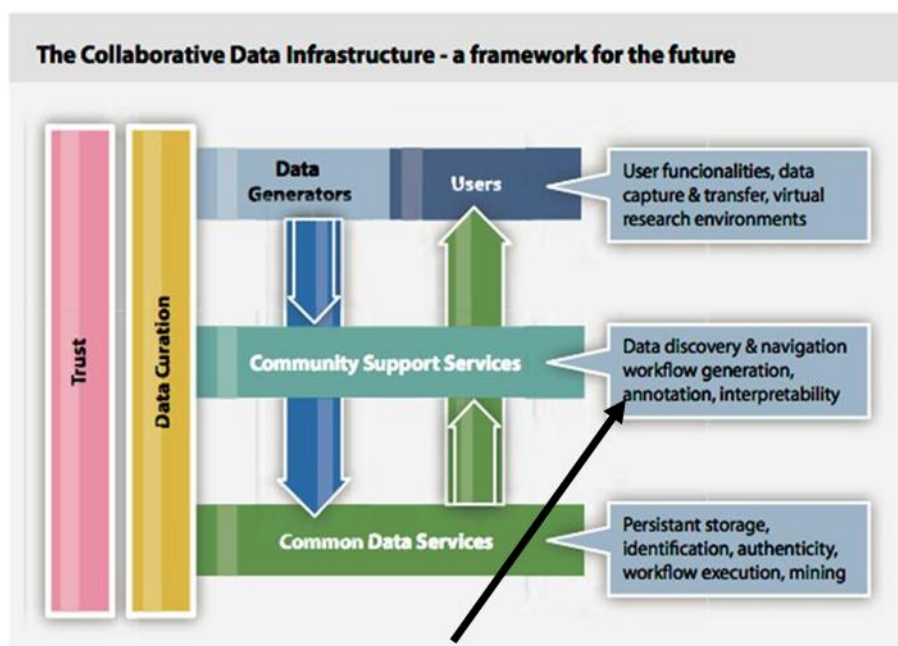
The overarching question in the context of APARSEN then is: *What*, beyond the primary data itself, must be preserved to retain richness of expression, in particular in terms of quality and provision and preservation of reputation.

### 1.1 Annotation

In many cases explanations/annotations necessary to understand and reuse a particular set of research data cannot be provided in a standardised form. This may result from a lack of knowledge concerning appropriate standardised metadata schemas. There is also a possibility that no suitable metadata schema exists. Obviously it is desirable to establish more standardized forms of data annotation in order to

- a) facilitate reuse – especially across subject boundaries,
- b) help streamline the process of annotating and
- c) reduce the complexities of long term data preservation.

But certainly there will always be cases where information closely related to a dataset cannot be encoded in metadata, as evident from the examples provided in section 2.1 Annotation and annotation services. It was therefore decided within APARSEN to follow the example of the High Level Expert Group on Scientific Data and use the term annotation instead of metadata, as in illustration 1.



**Illustration 1 Collaborative Data Infrastructure<sup>1</sup>**

In everyday linguistic usage, an annotation is a note that a person makes while reading any form of text. Annotation can in the simplest form be just underlining or highlighting to mark important words or passages, or adding explanatory comments or assessments to certain passages.

These annotations to text may be created for the private purpose of the reader, e.g. to quickly identify important parts of the text on subsequent reading. Annotations can also be shared between several users, e.g. in collaborative writing, editing or commenting. Three different aspects of annotation are identified in Agosti et al.<sup>2</sup>:

- comprehension and study: annotating a text is used to investigate and understand a concept better. These annotations are mostly private, as the consumer of the annotation is the creator. It is noted though, that other people reading an annotated text could as well benefit from existing annotations.
- interpretation and divulgation: annotations are used to comment and explain a text, to make it more comprehensible or to exchange ideas on a topic, such as an expert in literature annotation a demanding piece of art. Here, annotations are intended to be public, i.e. the consumers are people other than the creator.
- cooperation and revision: annotations are used to review someone else's work, i.e. annotations are a way to share ideas and opinions to improve the original work. Here, annotations are used collectively.

<sup>1</sup> High Level Expert Group on Scientific Data (2010). Riding the wave. How Europe can gain from the rising tide of scientific data. Final report, European Union, 31. <http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf>

<sup>2</sup> Agosti, M., Ferro, N. Annotations: Enriching a digital library. (2003). Research and Advanced Technology for Digital Libraries, Volume 2769 of Lecture Notes in Computer Science, Springer



The latter forms of annotation have seen a significant use with the emergence of digital distribution and publishing forms, when the annotations to a text can easily be shared with a larger user base. Thus, annotations are a topic prominent in digital library research.<sup>3</sup>

In the context of the APARSEN project, annotation is understood as “information added to data and it may be argued that all kinds of metadata are special types of annotation”<sup>4</sup>. This is a definition similar to the one presented by e.g. Shabajee et al.<sup>5</sup>: “we define annotation as metadata created after the creation of the content. It is this post hoc nature (“a note added to anything written”, Oxford English Dictionary, 1998) that represents a considerable expansion of its usefulness, as a means of adding value to content, because it now allows people other than the original content author to add metadata descriptions.”

These definitions and deliberations on annotation have their roots in text and text-related (e.g. facsimile) content. In recent years however, the concept has been transferred to mean also information added to data. The most prominent use is in genome annotation<sup>6</sup>, where GenBank, the data repository for nucleotide sequence, defines<sup>7</sup>:

“Feature annotation is the addition of biological features such as genes and associated coding regions, structural RNA, variation information, exon, introns, etc. to your submitted sequence.” In other words: Annotation here adds meaning to an otherwise meaningless sequence of molecule “names”. GenBank observes that “Adding feature annotation will also frequently provide an additional tool for reviewing the quality of primary nucleotide sequence data.” and “...annotating protein-coding regions will frequently highlight potential errors in the nucleotide sequence, such as insertion/deletions (in/dels) or improper or uncertain base calls that result from the sequencing reads.”

Usually, this annotated sequence entry will also link back at the original article which describes it and, in particular, the method of generating the annotation. These methods, in turn, can be expressed as workflows, see section 1.3.1.2, employing large number (e.g. 20) of tools to arrive at just a single annotation. Since the field of genomics and bioinformatics is in fast development, it can happen that at the time of review of the article, one of the tools is no longer available in the version used and reviewers require authors to re-do their annotation with the available, newer version of that tool. In summary, GenBank relies on articles to provide documentation and journals can do no more (at this time) than require reproducibility *at the time of review*.

We will subsequently discuss both user-generated and automatically generated annotations.

---

<sup>3</sup> Marshall, C. C. Annotation: from paper books to the digital library. (1997). In Proceedings of the second ACM international conference on Digital libraries, ACM.

Agosti, M., Ferro, N. Annotations: Enriching a digital library. (2003). Research and Advanced Technology for Digital Libraries, Volume 2769 of Lecture Notes in Computer Science, Springer

Gazan, R., Social annotations in digital library collections. (2008). D-Lib Magazine, 14(11/12)

Arko, R., Ginger, K., Kastens, K., Weatherley, J. (2006), Using annotations to add value to a digital library for education. D-Lib Magazine, 12(5)

<sup>4</sup> Description of work package 26 within the APARSEN Application, p. 47.

<sup>5</sup> Shabajee, P., Miller, L. (2002). Adding value to large multimedia collections through annotation technologies and tools: Serving communities of interest. In Proceedings of the Museums and the Web Conference.

<sup>6</sup> Stein, L.; Genome annotation: from sequence to biology; Nature Reviews Genetics 2, 493-503, doi:10.1038/35080529

<sup>7</sup> Annotating your Sequence for Submission - The GenBank Submissions Handbook.  
[http://www.ncbi.nlm.nih.gov/books/NBK53711/#gbankquickstart.what\\_do\\_you\\_mean\\_by\\_feat](http://www.ncbi.nlm.nih.gov/books/NBK53711/#gbankquickstart.what_do_you_mean_by_feat)

### 1.1.1 User-generated annotations

Many of today's Digital Library systems incorporate some sort of annotation tool that allows the user to create personal or shared annotations to the documents available in the collection.

An early investigation on how these tools could be implemented was performed in Marshall et al.<sup>8</sup>, starting with a study on how university students annotate their text books. The study draws on a few implications for annotation tools in digital library systems, such as to allow for "in situ annotation, distinguishable from the source", allow for well-known annotation forms such as underlining, highlighting and other, individual forms of annotations, or "smooth transitions between public and private annotations".

Marshall et al.<sup>9</sup> focus on the latter aspect. It is noted that on paper, sharing of personal annotations is often not intentional, e.g. when photo-copying materials that were annotated by the original owner/user, and that a huge percentage of annotations made in digital documents are personal, and not shared. Gazan<sup>10</sup> argues that another way in which initially private annotations are shared is e.g. when students buy used text books, which have been annotated by the previous owner(s). The author further argue that these annotations, even though the reason why they were created may be unknown, attracts more attention from the reader, and thus can help in making learning less of a solitary effort.

Agosti<sup>11</sup> identifies a number of annotation signs and discussed their implementation in the OpenDLib digital library system. Arko et al.<sup>12</sup> discuss how annotations in a digital library for education can be improved to foster student collaboration and understanding of subjects. It is noted that one shortcoming is the lack of a common framework for these annotations, and therefore proposes a framework for metadata records, of which annotation metadata is a key service. The metadata records are based on the architecture in the National Science Digital Library (NSDL).

Gazan<sup>13</sup> also touches issues of authority of (unknown) annotators that inform, challenge and often confuse subsequent readers. While the trustworthiness of these annotators might be questionable at first thought, at least the fact that they took the time to read and annotate the text, which gives the subsequent reader additional perspectives from which to evaluate the usefulness of the text. This is somewhat similar to social computing in the Web 2.0, where digital objects are most often also annotated by peers (comments, ratings, ...), rather than experts. Aggregate peer authority, i.e. the concordance of multiple peers, can become a source of trust. The author argues that Digital Library system have to embrace such approaches for their systems.

A storage solution for annotations, specifically for semantic web, was developed in the Annotea project of the W3 Consortium<sup>14</sup>. An RDF based format is defined, which lets users easily create, merge and mix annotations with other metadata. This metadata can then be stored locally, or in dedicated

---

<sup>8</sup> Marshall, C. C. Annotation: from paper books to the digital library. (1997). In Proceedings of the second ACM international conference on Digital libraries, ACM.

<sup>9</sup> Marshall, C. C., Brush, A. B. From personal to shared annotations. (2002). In CHI '02 extended abstracts on Human factors in computing systems, ACM.

<sup>10</sup> Gazan, R., Social annotations in digital library collections. (2008).D-Lib Magazine, 14(11/12)

<sup>11</sup> Agosti, M., Ferro, N. Annotations: Enriching a digital library. (2003). Research and Advanced Technology for Digital Libraries, Volume 2769 of Lecture Notes in Computer Science, Springer

<sup>12</sup> Arko, R., Ginger, K., Kastens, K., Weatherley, J. (2006), Using annotations to add value to a digital library for education. D-Lib Magazine, 12(5)

<sup>13</sup> Gazan, R., Social annotations in digital library collections. (2008).D-Lib Magazine, 14(11/12)

<sup>14</sup> Koivunen, M.(2005). Annotea and semantic web supported collaboration. Workshop on User Aspects of the Semantic Web (User-SWeb) at European Semantic Web Conference

Annotea servers, so that they can also be utilised by other users. The Annotea project also provided an editor for annotations, Amaya<sup>15</sup>; another implementation is provided as a browser-extension for Firefox<sup>16</sup>.

## 1.1.2 Automatic annotation

In several information retrieval domains, automatic annotation is a concept of automatically attaching metadata to objects. One prominent example is automatic image annotation, where the goal is to automatically assign a set of captions and descriptive keywords describing the content of the image<sup>17</sup>. Often approaches to this task are by a statistical machine learning model that can predict a set of annotations for an image given a training set of images already annotated. The learning is based on characteristic features extracted from the images. Automatic image annotation has the advantage that users can query for images using the natural-language keywords, rather than having to specify abstract image properties such as colours or texture. Similar approaches exist also for other types of media.

Agosti et al.<sup>18</sup> also suggest that textual data inside a digital library could be automatically annotated by the topics (categories) being assigned to certain subsections of the document, and that the document can then be organised not only by its original structure, but segmented into these topics.

Data provide an even wider potential to derive annotation: In the case of geospatial data, e.g. it is quite useful to associate place names with geographical co-ordinates, or vice versa. This “cross-walk” function is mediated by so-called gazetteers<sup>19</sup>. Since place names can change over time (or actually places such as islands can even move over long times<sup>20</sup>) or are actually challenged due to political reasons (Malvinas vs. Falkland Islands; Taiwan vs. Republic of China) it will be a non-trivial task to decide whether and how to provide the cross-walk: “Just in time” on search at a portal or search engine or as a persistent annotation in the repository?

Similarly, names of parameters measured and names and classification of objects observed will not be unambiguous: Whether is advisable or even possible to use a modern name for a parameter when the ancient one also indicates a different method of measurement or “translating” a species name to a “better” one, according to modern interpretation, is clearly questionable, as can be seen from the many names and classification for whales<sup>21</sup> and other species, currently known by the Encyclopaedia of Life.

In any case, re-assignment of such annotations even if possible through a simple database call, may sometimes be of major scientific relevance, should not be done lightly and the original annotation will probably need to be preserved as a matter of authenticity and provenance.

---

<sup>15</sup> <http://www.w3.org/Amaya>

<sup>16</sup> <http://annozilla.mozdev.org>

<sup>17</sup> Li, J., Wang, J.Z. (2008) Real-time computerized annotation of pictures. IEEE Trans. Pattern Analysis and Machine Intelligence (TPAMI). 30(6)

<sup>18</sup> Agosti, M., Ferro, N. Annotations: Enriching a digital library. (2003). Research and Advanced Technology for Digital Libraries, Volume 2769 of Lecture Notes in Computer Science, Springer

<sup>19</sup> Wikipedia: List of gazetteers [http://en.wikipedia.org/wiki/Gazetteer#List\\_of\\_gazetteers](http://en.wikipedia.org/wiki/Gazetteer#List_of_gazetteers)

<sup>20</sup> The island “Trischen”, formerly known as “Buschsand” moves at 30 m per year - <http://de.wikipedia.org/wiki/Trischen>

<sup>21</sup> Encyclopedia of Life: Cetacea - Dolphins, Porpoises, And Whales, Names, <http://eol.org/pages/7649/names>

### 1.1.3 Metadata in Digital Preservation

For digital preservation, metadata is of particular importance, as it is information that supports and documents the digital preservation process. Woodyard<sup>22</sup> identifies four tasks for preservation metadata:

- List the technical details about files and structure of the resource and how to use it.
- Record the history of all actions performed on the resource, including any changes or decisions made about it.
- Prove the authenticity through technical means and account for the continued custody of the resource.
- Retain information on who has the responsibility and rights to perform preservation actions on the resource.

There are several tools that generate automatic annotations for files. Format identification (and validation) can be performed by tools such as JHOVE<sup>23</sup>, Droid<sup>24</sup> and the UDFR (Unified Digital Format Registry)<sup>25</sup>. Extraction of metadata from files is supported by the National Library of New Zealand Metadata Extraction Tool<sup>26</sup>.

The PREMIS (Preservation Metadata: Implementation Strategies) working group has published the PREMIS data dictionary<sup>27</sup>, which defines a number of metadata fields that should be used for describing digital objects.

### 1.1.4 References and further reading

Agosti, M., Ferro, N. Annotations: Enriching a digital library. (2003). Research and Advanced Technology for Digital Libraries, Volume 2769 of Lecture Notes in Computer Science, Springer

Arko, R., Ginger, K., Kastens, K., Weatherley, J. (2006), Using annotations to add value to a digital library for education. D-Lib Magazine, 12(5)

Gazan, R., Social annotations in digital library collections. (2008).D-Lib Magazine, 14(11/12)

Koivunen, M.(2005). Annotea and semantic web supported collaboration. Workshop on User Aspects of the Semantic Web (User-SWeb) at European Semantic Web Conference

Li, J., Wang, J. Z. (2008) Real-time computerized annotation of pictures. IEEE Trans. Pattern Analysis and Machine Intelligence (TPAMI). 30(6)

Marshall,C. C. Annotation: from paper books to the digital library. (1997). In Proceedings of the second ACM international conference on Digital libraries, ACM.

Marshall,C. C., Brush, A. B. From personal to shared annotations. (2002). In CHI '02 extended abstracts on Human factors in computing systems, ACM.

---

<sup>22</sup> Woodyard, D. (2002). Metadata and preservation. Information services & use, 22(2).

<sup>23</sup> <http://hul.harvard.edu/jhove>

<sup>24</sup> <http://www.nationalarchives.gov.uk/PRONOM/Default.aspx>

<sup>25</sup> <http://www.udfr.org>

<sup>26</sup> <http://meta-extractor.sourceforge.net/>

<sup>27</sup> PREMIS Editorial Committee. (2008). Premis data dictionary for preservation metadata. Technical report, March.

PREMIS Editorial Committee. (2008). Premis data dictionary for preservation metadata. Technical report, March.

Shabajee, P., Miller, L. (2002). Adding value to large multimedia collections through annotation technologies and tools: Serving communities of interest. In Proceedings of the Museums and the Web Conference.

Woodyard, D. (2002). Metadata and preservation. Information services & use, 22(2).

## 1.2 Reputation

Preparing research data for proper preservation and reuse entails effort and costs. The willingness of researchers to bear this burden is closely linked to associated rewards. Increasingly, important scientific bodies assert not only that indeed rewards are due but also that this act is a scholarly achievement of its own.<sup>28</sup>

Consequently, one approach to enable rewards for data publication is to add and maintain functions to research data repositories to capture and preserve information which associates scientific reputation with individual data sets.

Once reputation for research data sets becomes an accepted currency in the competition among scientists, then research data repositories will to some extent have to act like journals. In this context this refers first to the dissemination activities typically associated with research journals. This will immediately lead to a desire to enhance the reputation of the data repository, respectively of the data sets available via the repository, through e.g. selectivity or else, explicit or implicit ranking.

Selectivity in journals is meant to transfer reputation built by a journal to each of the items published from then on. This mechanism will to some extent be made use of by data repositories as well. Ranking within the holdings of one repository does establish reputation on a scale calibrated by the repository through the process employed, in particular through pre-screening of the items submitted and selectivity regarding the clientele allowed to submit judgments.

As mentioned in various places within this report, and taken up here again: Annotation, reputation and data quality are closely interlinked. In this section a fourth category, visibility, needs to be made explicit because reputation cannot be achieved without it.

### 1.2.1 Citation, persistent identifiers and research data repositories

Once the quality or ranking of a dataset is established, the further build-up of its reputation and – even more important – the transfer of this reputation to the reputation of its creator is dependent on the ability to cite data sets, unambiguously and persistently.

---

<sup>28</sup> Most recently in June and July 2012: „Science as an open enterprise“, The Royal Society Science Policy Centre report 02/12, <http://royalsociety.org/policy/projects/science-public-enterprise/report/> and „Empfehlungen zur Weiterentwicklung der Wissenschaftlichen Informationsinfrastrukturen in Deutschland bis 2020“, Wissenschaftsrat, Drs. 2359-12, Berlin <http://www.wissenschaftsrat.de/download/archiv/2359-12.pdf>

ODE-report D3.2 Baseline Report: [http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2011/11/ODE-WP3-DEL-0002-1\\_0\\_public\\_final.pdf](http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2011/11/ODE-WP3-DEL-0002-1_0_public_final.pdf)

ODE-report D4.1 Executive summary of Report of Integration on Data and Publication: [http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2011/11/ODE-ReportOnIntegrationOfDataAndPublications-exesummary\\_final.pdf](http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2011/11/ODE-ReportOnIntegrationOfDataAndPublications-exesummary_final.pdf)

In science, technology, medicine (STM) publishing practice, most articles are cited as a whole, while in the “book-sciences” it is clearly necessary to refer to individual pages, if not lines. In analogy, considering the potential size or complexity of data sets, it is desirable to provide the possibility to cite subsets. Beyond this there is a problem quite outside of reasonable analogy with journal articles: that of enhancements (read: correction of minor errors) and extensions (e.g., in coverage of space and time) of datasets. Should those retain the same citation? Should versioning be employed in one way or another? How to reflect this through identifiers?

These questions have been discussed, e.g. for an important corpus of economic data by Green<sup>29</sup> and by others<sup>30</sup>, but are still not being answered consistently and unanimously. In the case of enhancement and extensions, practical limitations are brought forward, such as by Green: “In the case of dynamic datasets the volume of change can be so large or frequent to make tracking back impossible to manage.”

It goes almost without saying that these functions can best be implemented – reliably and impartially – by professional research data repositories. Research data which are expected to gain reputation should therefore be made available via suitable data repositories, not individual web- or ftp-sites. The great majority of research data is not yet stored in such repositories. This may change drastically as cultural habits of research change – in particular, whether or not to cite datasets and to “count” them on an equal footing with articles for evaluations. For that change to actually happen and to take hold, the above questions need to be fully answered and adequate practices be implemented firmly and transparently by repositories.

## 1.2.2 Research data journals

Research data certainly do not stand on their own – in most cases they serve as the facts underlying journal articles of scholarly books. Today, if at all, they are referred to from these texts via links or haphazard wording within the text or they are provided as “supplemental information”, available from the online systems of publishers. The customary processes of quality assurance, such as peer review of articles, do not extend to these external sources or supplements, creating mounting concern.<sup>31</sup> One could extend these concerns about journals by saying that using data of uncertain reputation is beginning to undermine the reputation of journals themselves.

The new breed<sup>32</sup> of data journals is focused on the publication of original research data sets and corresponding explanations (e.g. Earth System Science Data (ESSD))<sup>33</sup>. These journals create an opportunity to gain reputation for the researcher by publishing high quality “annotations” to their high quality data.

---

<sup>29</sup> Green, T. (2009); “We Need Publishing Standards for Datasets and Data Tables“, OECD Publishing White Papers, OECD Publishing, doi:10.1787/787355886123

<sup>30</sup> Lawrence, B., Jones, C., Matthews, B., Pepler, S., Callaghan, S. “Citation and Peer Review of Data: Moving Towards Formal Data Publication”, International Journal of Digital Curation, 2011, Vol. 6, No. 2, pp. 4-37, doi:10.2218/ijdc.v6i2.205

<sup>31</sup> Maunsell, J., „Announcement Regarding Supplemental Material“, Journal of Neuroscience, 30(32):10599-10600, 2010. <http://www.jneurosci.org/cgi/content/full/30/32/10599>

<sup>32</sup> Overviews have been worked out in APARSEN WP33, “Report on Peer Review of Research Data in Scholarly Communication” <http://www.alliancepermanentaccess.org/index.php/knowledge-base/member-resources/documents-and-downloads/?did=82> and ODE WP4, Kotarski R, Reilly S, Schrimpf S, Smit E, Walshe K (2012). Report on best practices for citability of data and on evolving roles in scholarly communication”, <http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2012/08/ODE-ReportBestPracticesCitabilityDataEvolvingRolesScholarlyCommunication.pdf>

<sup>33</sup> <http://www.earth-system-science-data.net>



There is a clear distinction between classical "article journals", data repositories and the new "data journals". Data journals not only focus exclusively on the making available of research data, they also provide quality assurance mechanisms specifically geared towards research data. The journal mentioned above, *Earth System Science Data* (ESSD), is an example for such a policy.

"The [journals] peer-review secures that the data sets:

- are at least plausible and contain no detectable problems;
- are of sufficiently quality and their limitations are clearly stated;
- are open accessible (toll free), well annotated by standard metadata (e.g., ISO 19115) and available from a certified data center/repository;
- are customary with regard to their format(s) and/or access protocol, however not proprietary ones (e.g., Open Geospatial Consortium standards), expected to be useable for the foreseeable future."<sup>34</sup>

By thus "wrapping" a peer reviewed article around data, it becomes a matter of obeying good scientific practise to cite it. Journals exclusively devoted to data publishing avoid the danger of "data briefs" in classical journals or similar means, which might let data publications appear second class citizens in scholarly publishing, diminishing their reputation. This function may help in the interim, until data citation is a matter of course. In the following section on quality it will be made clear that peer review is not applicable to all data, so the other function of data journals, providing reputation to creators, is not available to all.

An interesting lemma arising for preservation is the necessity to establish and preserve a bilateral link between the dataset in its repository and the data journal article (perhaps technically declared as an annotation to the dataset).

### 1.2.3 Commenting and recommendation functions

Commenting or recommending data in repositories is another approach of adding reputation to data sets. The data review functionality that is part of the EASY system developed and implemented by the Dutch Data Archiving and Networked Services (DANS) is an example. EASY enables users to appraise data sets by awarding one to four stars.<sup>35</sup> This and similar examples are obviously modelled after the successful and useful precedent of Amazon and similar commercial entities.

An advantage of the review function is that the appraisals can be measured easily. The simplicity of the system also entails a disadvantage. The system does not record the reasons for the appraisals. This problem is well known from classical articles: A high citation rate is normally assumed to indicate high quality and relevance of the article. This measurement does not however record the semantic context of the citations. The article might as well get a lot of attention for a negative reason, e.g. bad practice or erroneous conclusions (The article about "cold fusion" was one of the most-cited articles ever in physics).

Whatever the mechanism and its measurements, its results will need to be preserved along with the data itself, to preserve reputation. An interesting migration challenge might arise when a repository changes its mechanisms or when data need to be transferred (handed on) or replicated to other repositories.

---

<sup>34</sup> [http://www.earth-system-science-data.net/general\\_information/about\\_this\\_journal.html](http://www.earth-system-science-data.net/general_information/about_this_journal.html) (14 August 2012)

<sup>35</sup> See section 2.2.3 Example: EASY (DANS).

## 1.2.4 Visibility

The building-up of reputation is closely linked to visibility. Therefore, research data repositories do not only have a function as providers of tools enabling users to appraise data sets. They can also increase their visibility. This is achieved via several avenues:

- Research data repositories make data sets visible to search engines.
- Data sets can be linked via annotations or metadata. Based on these links users of one data set can be guided to related data sets.
- The power of this referral function is greatly enhanced if a research data repository is cooperating with journals whose author deposit data related to their articles in this particular repository. In this way the audience of the journals become potential audience for the data repository thus raising the chance of being discovered for all data sets of this repository.
- Annotations may very well contain information which could be used to measure appraisals of a data set. In fact, annotations have the potential of being much more helpful to measure reputation than streamlined systems like the above described data review tool of EASY, because the user can explain what he thinks is important concerning a data set. The yet unsolved challenge is the automatic extraction of these hidden appraisals.

The perhaps boldest move with respect to visibility is the recent announcement of Thomson Reuters to create its “Data Citation Index”. Considering that in many cases journal articles are included in evaluations only if published in a journal on Thomson Reuters’ master journal list, Thomson’s selection criteria<sup>36</sup> will challenge repositories in many ways, in particular in the long term to make datasets not just citable, but take care that the number of citations is actually non-negligible.

## 1.3 Data quality

The term quality is related to fitness for a purpose. Conversely, if quality of data is (completely) unknown, the data are not fit for any purpose of re-use. Therefore, it may not be justified (except perhaps for historical reasons) to preserve unqualified data. Were today’s repositories to be judged by this argument, it might turn out that much if not most of their contents is in dire straits.

Whatever the practice to create, measure, assure or otherwise determine quality, it will entail creating evidence which practices have been applied by whom and when and, of course, an estimate of error. Both evidence and estimate need to go along with the data itself and be preserved with it – the former perhaps for a limited time, after which no further scrutiny is to be expected, the later for as long as the data itself, because it will be needed on re-use of the data, in calculations of error propagation.

Today, we see (e.g. in the examples in chapter 2) these principles realized fully for very few data. However, the topic of data quality is experiencing an explosive growth of attention, from funding bodies<sup>37</sup> and scientific societies<sup>38</sup> to data practitioners<sup>39</sup>. Consequently, the means to preserve quality-related information in a structured way will be in high demand, soon.

---

<sup>36</sup> „Repository evaluation, selection and coverage policies for the data citation index within Reuters web of knowledge” (2012) [http://wokinfo.com/media/pdf/DCI\\_selection\\_essay.pdf](http://wokinfo.com/media/pdf/DCI_selection_essay.pdf)

<sup>37</sup> EUROHORCS-ESF Task Force, “EUROHORCS and ESF Vision on a Globally Competitive ERA and their Road Map for Actions”, 2009. [http://www.eurohorcs.org/SiteCollectionDocuments/ESF\\_Road%20Map\\_long\\_0907.pdf](http://www.eurohorcs.org/SiteCollectionDocuments/ESF_Road%20Map_long_0907.pdf)

<sup>38</sup> „Science as an open enterprise“, The Royal Society Science Policy Centre report 02/12, <http://royalsociety.org/policy/projects/science-public-enterprise/report/>

<sup>39</sup> Robinson, E., Meyer, C.B., Lenhardt, W.C., “Moving the science data quality dialogue forward”, EOS, Transactions



As to the practices or processes to establish quality two essentially different approaches can be identified<sup>40</sup>:

*“If ... a potentially large number of articles are expected to rely on a (comprehensive or exceptional) dataset — also known as re-use of data — there is no way around the need to make sure, as far as possible, that this dataset itself is reliable. There may be communities of practise, e.g., in remote sensing or monitoring of environmental data, which work by established practises of documenting, testing and calibration of instruments, complemented by methods of (semi-)automatic validation of results. As long as those instruments and methods are operated by experienced, professional staff it may suffice for quality assurance to affirm just that, and by making all necessary documentation available. One might think of a priori quality assurance, here.*

*However, especially in pure research, there are many innovative and evolving and therefore less thoroughly documented and tested methods, which nevertheless produce substantial results, i.e., valuable data. It is this subset, which needs to be subject to quality assurance a posteriori. How can this, to put it loosely, "quality assessment with somewhat incomplete and/or ingenious documentation/proof" be done? One "obvious" answer is: Peer review, a method already practised and reasonably well understood by the parties involved.”*

Were dataset just to be seen as stand-alone items, both approaches extend metadata or data themselves by entries for error estimates and create “auxiliary” documents and other evidence. These items can be treated as it has already been discussed in the more general sections on annotation and reputation.

The Royal Society report „Science as an open enterprise” begins with the words: *“Open inquiry is at the heart of the scientific enterprise. Publication of scientific theories - and of the experimental and observational data on which they are based - permits others to identify errors, to support, reject or refine theories and to reuse data for further understanding and knowledge. Science’s powerful capacity for self-correction comes from this openness to scrutiny and challenge.”*<sup>41</sup> These reflections challenges us to widen the field of view to include the research and publication processes as a whole, to include to processes of reception and scrutiny and its technical means in the presence of huge amounts or complex data.

The following sections provide an array of current thinking into the merging of data and computation into the research and publication process as it influences and provides quality and describe or at least hint at some of the challenges emerging for preservation.

### 1.3.1 Interweaving publications and research data

Most recent research is based upon data that also exists in digital format or even is “born digital” and is never printed. The validity and coherence of research data is a crucial factor determining the quality of experiments and the resulting publications. Considering the flood (“deluge”) of data, publications can no longer be seen as standalone resources that include all necessary properties and information about the research topic anymore. The data collected and analysed during the research process belong to the publication as well and have therefore to be reviewed in their own right. Tools that support peer review processes of publications and data increase the data quality immediately. This awareness

---

American Geophysical Union, (2012) Vol. 93, No. 19, p. 189, doi:10.1029/2012EO190008

<sup>40</sup> Pfeiffenberger, H., Carlson, D., ““Earth System Science Data” (ESSD) — A Peer Reviewed Journal for Publication of Data”, D-Lib Magazine (2011) Vol 17, No.1/2, doi:10.1045/january2011-pfeiffenberger

<sup>41</sup> The Royal Society (VI.2012): Science as an open enterprise, Science Policy Centre report, 02/12, London, The Royal Society. [royalsociety.org/uploadedFiles/Royal\\_Society\\_Content/policy/projects/sape/2012-06-20-SAOE.pdf](http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE.pdf)

fosters the quality of data as it encourages scientists to deliver highly maintained data sets. A further concept is the close combination of research publications and the data they are based on. There exist different ideas for allowing researchers to rerun experiments and analyse the claims the original research stated. So-called executable papers integrate research data directly into the corresponding publications. This allows other researchers to conveniently reuse data, retrace the scientific workflow, or verify results, thus enhancing the quality of data by inherent quality control. Scientific workflow management systems (SWMS) allow the reproduction of complex experiments in a broader context, as they allow researchers to model their experiments in a formal way. SWMS further allow automated runs of experiments and therefore introduce exact reproduction in defined environments. This further enhances the data quality and can easily be integrated into existing workflows.

The following section gives an introduction to executable papers, which allow researchers to rerun experiments that lead to a publication and therefore validate and test the hypothesis under examination.

### 1.3.1.1 Peer review and reproducibility

Research data have to meet stringent data quality standards in order to be useful to the research community. Data quality can be achieved by many different factors, some of which have already been introduced in this section. Additional to technical instruments and methods to improve the quality of data, the concept of reviewing is an essential tool for assessing the quality of research. A fundamental requirement for reviewing scientific experiments and theories is the possibility of reproducing the statements made by scientists in their publications. The validity of an experiment can only be judged if it is possible to rerun a specific experimental setup under the same preconditions. This essential standard applies to digital data more than ever, as the majority of research data is already digitally available. The need for replicating scientific experiments by providing the underlying data has also been addressed by governments who encourage scientists to make their data available to peers<sup>42</sup>. But data is only verifiable if it is accessible, understandable and coherent. This has to be ensured by using standardized data formats and by checking for compliance with best practices of the field. The quality of data can be enhanced further if reviewers are supported by tools enabling them to rerun experiments and verify research results. The easier the access, usage and application of data and the corresponding methods that were used, the lower are barriers for reviewing personnel and readers.

Peer review has a long tradition in research as the number one tool for increasing the quality of research by assessing the contributions made by fellow researchers. The same is valid for data quality, although it is even more difficult to judge as this often requires very specialized knowledge. Using standard data formats increases the general data quality, but manual work is still required. Therefore tool support is an essential factor for decreasing barriers for assessing and reusing research data.

---

<sup>42</sup> House of Commons; Science and Technology Committee (28 July 2011): Peer review in scientific publications, Eighth Report of Session 2010-12, HC856, London: The Stationery Office Limited, <http://www.publications.parliament.uk/pa/cm201012/cmsselect/cmsctech/856/85602.htm>

### 1.3.1.2 Enhanced publications, executable papers and rich internet publications

There exist different approaches that allow the integration of research data into publications in order to enhance the quality of the overall outcome. Concepts such as enhanced papers<sup>43</sup> and rich Internet publications<sup>44</sup> enhance the understandability of publications by augmenting them with supplementary content. Enhanced papers are publications that are augmented with links to additional multimedia content, such as full-text articles, comments, images and other sources available online and also to research data. Rich Internet publications feature multimedia content and interactive elements that support the visualization of research results, such as interactive maps or tools for data analysis. Both concepts are valuable, but do not allow rerunning experiments on original data. This issue is tackled by executable papers<sup>45</sup> which allow executing and therefore rerunning scientific workflows. These are especially useful to increase data quality as they enable peers to detect quality problems regarding the research data. Connecting publications closer with their constituting data enhances the validation of experiments and fosters the reuse of this data. Both factors influence the quality of research in a positive way. An overview of these approaches is given in the APARSEN “Report on peer review of research data in scholarly communication”<sup>46</sup>.

### 1.3.1.3 Problems, challenges and chances

Reviewing digital research data raises several new questions and problems in comparison with traditional paper-based peer reviews of scientific work. The responsibility for the data quality is distributed amongst different parties.

From the researcher’s perspective, submitting research data along with the actual paper or report has positive effects on the data quality itself and the overall quality of the research outcome. The awareness that their research data is peer reviewed imposes even more accuracy and carefulness in the preparation of the data. It promotes the proper preparation of the data for the review process and thus enhances reusability. This in turn fosters peers and scientists from related research areas to engage in verification and also falsification of results, as it is easier to rerun experiments on the very same data. By the same token, these advantages are associated with higher costs. Researchers need more time and effort in order to prepare the data for practical reuse. This includes the delivery of data in the proper formats and requires comprehensive documentation of the data sets. As the experimental data might require considerable storage space the distribution of the data sets can also become an issue. Intellectual property rights linked to research data pose another challenge. They have to be respected and treated in a way that complies to the requirements of researchers, reviewers and publishers. If the research results are to be publicly available – as sometimes demanded by funders –, the consideration

---

<sup>43</sup> Sierman, B., Schmidt, B., Ludwig, J. (2009) Enhanced Publications : Linking Publications and Research Data in Digital Repositories. Surf EU-Driver. Amsterdam University Press, Amsterdam.

<sup>44</sup> Breure, L., Voorbij, H., Hoogerwerf, M. (2011) Rich internet publications: Show what you tell. Journal Of Digital Information, Vol 12 (Nr. 1).

<sup>45</sup> Elsevier (14. Dec. 2010) Elsevier Launches Executable Paper Grand Challenge, press release, [http://www.elsevier.com/wps/find/authored\\_newsitem.cws\\_home/companynews05\\_01788](http://www.elsevier.com/wps/find/authored_newsitem.cws_home/companynews05_01788)  
David, K., Santos, E., Mates, P., Vo, H.T., Bonnet, P., Bauer, B., Surer, B., Troyer, M., Williams, D., Tohline, J., Freire, J., Silva, C. (2011). A provenance-based infrastructure to support the life cycle of executable papers. Proceedings of the International Conference on Computational Science, ICCS 2011

<sup>46</sup> Pampel, H., Pfeiffenberger, H., Schäfer, A., Smit, E., Pröll, S., & Bruch, C. (2012). Report on Peer Review of Research Data in Scholarly Communication. Retrieved from <http://epic.awi.de/30353/>

of property rights becomes even more obvious. This also entails that provenance information has to be generated and stored in order to be able to trace the custody and authorship of data. The same is valid for authenticity and integrity information of data. Authenticity and integrity refer to the completeness, validity and correctness of data.<sup>47</sup> It has to be ensured that transmitted experimental data are protected against deliberate or accidental manipulation. Privacy protection is another legal issue that regularly needs close attention when dealing with certain kinds of research data.

From the reviewer's perspective the complexity of scientific data is a major problem because assessing the quality and validity of data often requires domain specific knowledge for correct interpretation. This special know-how is often limited to a very small designated community. Substantial documentation can help to enhance the understanding of research data but it does not fully solve the problem of complexity. Understanding the relationships among data in a set requires deep analysis by the reviewers. Comprehensive data delivery puts another burden on the reviewers, as they have to take complex data analysis into account.

Many journals require blind or double blind peer review processes. This paradigm is essential for unbiased research and it needs to be applied to research data in the same manner as it is applied to publications. It has to be ensured research data remains uniquely identifiable. This requires the usage of persistent identifiers<sup>48</sup> and mechanisms to discover and access research data. On the positive side, having access to research data increases the insights which reviewers can gain. This dramatically enhances their ability to judge the contribution of some research to a specific research area. In many cases it also allows rerunning experiments, verification of the output and check the underlying data pool for consistency. This facilitates the discovery of errors in the data basis and the detection of sugar coated data with a much higher confidence.

Publishers can provide access to original research data to their consumers and customers. From the publishers' perspective, this constitutes a new service which can possibly be exploited commercially. The direct linkage between the research outcome and the underlying data grounding enables enhanced peer review capabilities that go beyond simple downloads of research data. Publishers also need to adapt their infrastructure if they want to provide (sustainable access to) comprehensive research data in addition to traditional papers.

As described in the APARSEN-“Report on Peer Review of Research Data in Scholarly Communication”<sup>49</sup>, repositories are responsible for receiving the researcher's data output, storing the data sets in adequate formats and perform quality checks on these data. Another important task is their long term preservation in order to keep them accessible for future research. Preservation is aggravated as research data are provided in many different formats which are often combined to complex research objects.

### 1.3.2 Reproducibility by using scientific workflow management systems

E-Science as a relatively new scientific paradigm has become more and more important in many fields of research. Most of modern experiments are simulated in computational environments and involve various steps for their execution. These *in silico* experiments are often denoted as scientific workflows.

---

<sup>47</sup> See APARSEN-Report "Report on Authenticity and Plan for Interoperable Authenticity Evaluation System" (2012), version 2.4. [http://aparsen.digitalpreservation.eu/pub/Main/ApanWp24/APARSEN-REP-D24\\_1-01-2\\_4.pdf](http://aparsen.digitalpreservation.eu/pub/Main/ApanWp24/APARSEN-REP-D24_1-01-2_4.pdf)

<sup>48</sup> See APARSEN-Report "Persistent Identifiers Interoperability Framework" (2012) [http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2012/04/APARSEN-REP-D22\\_1-01-1\\_8.pdf](http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2012/04/APARSEN-REP-D22_1-01-1_8.pdf)

<sup>49</sup> Pampel, H., Pfeifferberger, H., Schäfer, A., Smit, E., Pröll, S., & Bruch, C. (2012). Report on Peer Review of Research Data in Scholarly Communication. Retrieved from <http://epic.awi.de/30353/>

Scientific workflow is defined as the computer-assisted or automated execution of a scientific process, in whole or part, which usually streamlines a collection of scientific tasks with data channels and dataflow constructs to automate data computation and analysis to enable and accelerate scientific discovery. Such a workflow consists of various steps that require different computational processing units, the usage of various tools and the exchange between diverse systems. Specialized infrastructures and management tools are needed in order to orchestrate complex experiments. This class of applications is denoted as Scientific Workflow Management Systems (SWMS). Lin et al.<sup>50</sup> define scientific workflow management systems as "systems that completely define, modify, manage, monitor, and execute scientific workflows through the execution of scientific tasks whose execution order is driven by a computerized representation of the workflow logic". In contrast to Virtual Research Environments, which focus on collaboration between researchers and the sharing of computational resources<sup>51</sup>, SWMS highlight the workflow paradigm and the orchestration of services.

Taverna Workbench<sup>52</sup> is an open source project that allows to design and run workflows. It is a general purpose workflow engine, which can be used for various applications. Taverna allows to orchestrate various local and remote services and to model the data flow between its components in order to automate a process. Therefore it is widely used in the scientific community and used for modelling data centric experiments. It is written in the programming language Java<sup>53</sup> and distributed under the GNU Lesser General Public License (LGPL)<sup>54</sup>. Taverna provides a graphical user interface that allows scientists to design their experiments in a convenient way and to visualize the data flow of an experiment. An example of such a workflow is given in the figure below.

---

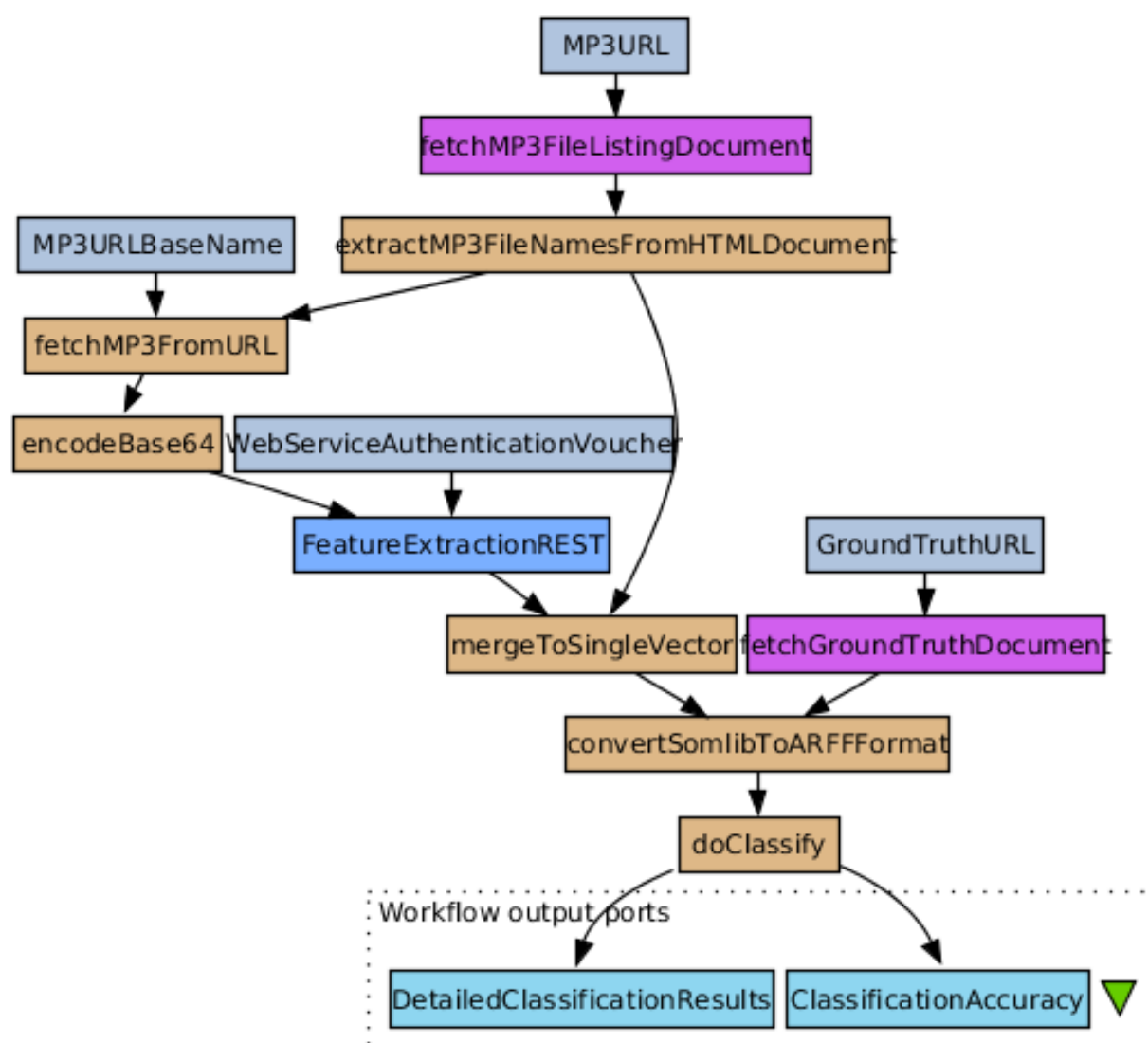
<sup>50</sup> Lin, C., Lu, S., Fei, X., Chebotko, A., Pai, D., Lai, Z., Fotouhi, F., Hua, J., (2009). A Reference Architecture for Scientific Workflow Management Systems and the VIEW SOA Solution. IEEE Transactions on Services Computing, 1(2), 79-92. doi 10.1109/TSC.2009.4

<sup>51</sup> <http://www.jisc.ac.uk/whatwedo/programmes/vre1>

<sup>52</sup> [www.taverna.org.uk](http://www.taverna.org.uk)

<sup>53</sup> [www.java.com](http://www.java.com)

<sup>54</sup> [www.gnu.org/licenses/lgpl.html](http://www.gnu.org/licenses/lgpl.html)



**Illustration 2 Scientific workflow modelled in the Taverna Workflow engine<sup>55</sup>**

This workflow is used to perform genre classification of music MP3s. It is modelled in Taverna by invoking different services (called processors) and performing various operations by providing the output of one step as input to further processing steps. Taverna provides various ready-to-use processors and can be extended by the use of the BeanShell<sup>56</sup> scripting language. Furthermore, Taverna can invoke remote services via REST interfaces and it provides a repository for Web services that can easily be integrated into a workflow. This allows the orchestration of complex scientific workflows. By using a SWMS like Taverna scientific experiments can be automated and therefore rerun in precisely the same manner as the original experiment has been performed. The representation of a workflow as a graph is convenient for describing the causal sequence of the experimental steps

<sup>55</sup> Mayer, R., Rauber, A., Neumann, M.A., Thomson, J., Antunes, G. (2012). Preserving scientific processes from design to publication. Proceedings of the 15th International Conference on Theory and Practice of Digital Libraries (TPDL 2012)

<sup>56</sup> [www.beanshell.org](http://www.beanshell.org)



and can also be used as a visualization tool for experiment design. Taverna logs details about the execution of a workflow in an internal database. This database contains provenance data, which can be used as additional metadata about the causal relationships between the invoked processors. This provenance data can be exported into different formats for further processing. Thus it is possible to store the data in the format of the Open Provenance Model, which is introduced in APARSEN-Report "Implementation and testing of an authenticity protocol on a specific domain" (2012).<sup>57</sup>

There exists a variety of SWMS for different purposes<sup>58</sup>, like Kepler<sup>59</sup> or VisTrails<sup>60</sup>. They have different backgrounds and stem from communities having other research backgrounds. These systems have in common that they enable easy reproduction of defined workflows in scientific domains. Executing, rerunning or validating scientific experiments becomes much easier for scientists using these systems. Using complex computational models and workflow engines does have many benefits for scientists and publishers, but introduces new challenges regarding the preservation of research experiments.

### 1.3.3 Long term preservation of scientific experiments

Scientific workflows as described in the previous section are used to model complex experiments in a detailed and formal way. Hence, their course of actions and data flows are specified and can be understood, interpreted and reproduced by peers for the verification of the results of the corresponding experiment. Such workflows are often assembled of many individual steps, which themselves rely on third party libraries and external resources, such as Web services.

Using resources that are beyond the area of influence of an individual scientist is a risk factor for the preservation of scientific experiments. However external components are essential for many scientific workflows. Although the data flow of an experiment might be clearly defined, remote resources can be considered as a black box component that reacts on provided inputs by delivering the desired result.

Their internal configuration or behaviour might be completely unknown and could be changed any time without prior information. This uncertainty impairs the reproducibility of research experiments whenever third party components or remote services are used<sup>61</sup>. The sample workflow described previously has several components that could cause problems because they rely on external resources. Web services can change their behaviour or become offline. Third party libraries can become obsolete as well and their maintenance can stop at any point in time. In contrast to Web services, local third party libraries have the advantage that they usually can be accessed and preserved for later reference.

In the workflow depicted in Illustration 2 the sample data is retrieved from a remote Web server, which provides a set of music files for testing the classification algorithm. In the next step, these sample files are transmitted to a remote Web service, which extracts the feature vector. This Web service is also located on a remote machine and it is only known that it accepts one MP3 file at once and that it requires an authentication voucher, which prevents the service from being abused by non-authorized users. In a parallel step, the so called ground truth, to which the extracted features will be

---

<sup>57</sup> [http://aparsen.digitalpreservation.eu/pub/Main/ApanWp24/APARSEN-REP-D24\\_2-01-2\\_2.docx](http://aparsen.digitalpreservation.eu/pub/Main/ApanWp24/APARSEN-REP-D24_2-01-2_2.docx)

<sup>58</sup> Curcin, V., Ghanem, M.. Scientific workflow systems - can one size fit it all?. (2008). Biomedical Engineering Conference, CIBEC 2008.

<sup>59</sup> [www.kepler-project.org](http://www.kepler-project.org)

<sup>60</sup> [www.vistrails.org](http://www.vistrails.org)

<sup>61</sup> De Roure, D., Belhajjame, K., Missier, P., Gomez-Perez, M., Palma, R., Ruiz, J. (2011) Towards the preservation of scientific workflows. 8th International Conference on Preservation of Digital Objects iPRES 2011.

compared to, is retrieved from a different URL on a different server. In the next step, a local but external library is used in order to classify the music features against the ground truth and return the result. From a preservation perspective, preserving the music files and the ground truth, which have been downloaded from remote servers, can be archived in a classical way. If these files exist in an archive, they can be provided for later reference to the workflow in order to obtain the same results. More difficult is the preservation of Web services, as they can disappear from the Web or change their behaviour at any point in time. Therefore precise descriptions of the requirements of the Web service results and their analysis are crucial. By using the information gathered in previous executions of a workflow, a simulation of a service becomes a solution to some extent. Also, previous responses from a remote service can be preserved and used as dummy-services, in order to understand – and if necessary, to redevelop – the operation of a workflow at a later point in time.

Third party libraries introduce more complexity to a workflow. These libraries might have dependencies on local execution environments, which also have to be preserved in order to perpetuate the compatibility with other components of a workflow system and vice versa. Therefore, detailed documentation of the versions of software libraries, operating systems and other components in use is necessary in order to preserve the SWMS and its processing components. This small sample workflow already demonstrates the complexity of the preservation of scientific workflows. Although SWMS add another layer of complexity and issues a challenge to the preservation community, they enhance data quality and the reproduction of experiments.

### **1.3.4 Conclusions and outlook of executable publications and scientific workflow management systems**

Today, the majority of scientific experiments is at least partly based upon computational steps and would therefore benefit from new forms of scientific publications closely connected to underlying data and software. Such a close connection enhances data quality of the whole scientific process. It allows peers to reproduce results and thereby enhances the quality of the data and the publication. The reason for this enhancement is on the one hand based on the fact that researchers have to publish their data sets. On the other hand this data has to go through a review process itself. Scientific workbenches allow design complex experiments without detailed knowledge about the systems themselves and they provide precisely defined workflow models without a lot of extra effort. The biggest benefit for the research community is the possibility of rerunning and thereby reproducing scientific experiments, which enhances the quality of research dramatically. In order to keep the insights gained for posterity, these experiments and the research data have to be preserved in archives, which can be accessed many years later. This is still a challenge, as the experiments get more and more complex and rely on external resources, which are beyond the scope of a single administrative entity.



## 2 Annotation, reputation and data quality: Approaches by APARSEN partners

This chapter illustrates approaches to the challenges annotation, reputation and data quality pursued by APARSEN partner organisations. Consultations with these partners were used as a basis to develop the questionnaire to inquire notions across APARSEN on these challenges.

### 2.1 Annotation and annotation services

The following examples point to some activities of APARSEN partners in the area of data annotation.<sup>62</sup> Two of the examples highlight why the broad term "annotation" was used instead of "metadata", one shows how far "classical" metadata can support research data preservation.

In the Netherlands the Alfalab project<sup>63</sup> aims at making different annotation systems interoperable. FORTH-ICS is developing software that helps update semantic descriptions.

#### 2.1.1 Example: Alfalab (DANS)

DANS<sup>64</sup> is partner in the Alfalab<sup>65</sup> project. This project is funded by The Royal Netherlands Academy of Arts and Sciences (KNAW) as part of its strategy of supporting humanities research in general, but in particular the digital (methods) and computational humanities. Digital methods within the humanities can be used to stimulate and promote interdisciplinary cooperation and synergy.

To stimulate and promote interdisciplinary cooperation, within the Alfalab project DANS is designing and developing a demonstrator that shows how heterogeneous annotations from different disciplines can be made interoperable using Open Annotation Collaboration (OAC)<sup>66</sup>, how these can be managed from a central point of access and how these can be interrelated to allow interdisciplinary discovery and/or usage of the data to create a synergy between disciplines.

OAC "seeks to facilitate the emergence of a Web and Resource-centric interoperable annotation environment that allows leveraging annotations across the boundaries of annotation clients, annotation servers, and content collections."<sup>67</sup>

OAC will be used to exchange annotations between the annotating tools and the central demonstrator. By using the OAC, and its fundamental linked data principles, DANS will demonstrate how heterogeneous annotations can be interoperable, how these can be interrelated and how these annotations can be related to other public data sources such as DBpedia, GeoNames, etc.

---

<sup>62</sup> The description based on <http://alfalablog.huygensinstituut.nl> and on <http://www.openannotation.org/wiki/index.php/User:AWitteveen>. [Version: 22 March 2011, 17:03]

<sup>63</sup> <http://alfalab.ehumanities.nl/>

<sup>64</sup> <http://www.dans.knaw.nl>

<sup>65</sup> <http://alfalab.ehumanities.nl/>

<sup>66</sup> <http://www.openannotation.org>

<sup>67</sup> [http://www.openannotation.org/wiki/index.php?title=Main\\_Page](http://www.openannotation.org/wiki/index.php?title=Main_Page). [Version: 9 August 2011, 15:39]

## 2.1.2 Example: Data Preservation in High Energy Physics" (DPHEP)

As data from high energy physics (HEP) experiments usually are complex and unique, the community started the initiative "Data Preservation in High Energy Physics" (DPHEP)<sup>68</sup> to preserve the output of the experiments (DPHEP, 2009 and DPHEP, 2012). They studied the complexity and diversity of the research data output in HEP. As there are many community standards and individual solutions within experiments, it was important to get an overview. The DPHEP group distinguished four different levels of research data in HEP with the most comprehensive layer including "basic level data" which is amongst others simulation/analysis software (DPHEP, 2009). This level is needed to maintain full potential of the experimental data. In the DPHEP model, data associated to a publication makes up the highest level of abstraction and therefore comprehension research data. This is what is handled on the data repository HEPData<sup>69</sup> and INSPIRE.

Preservation Model	Use case
1. Provide additional documentation	Publication-related information search
2. Preserve the data in a simplified format	Outreach, simple training analyses
3. Preserve the analysis level software and data format	Full scientific analysis based on existing reconstruction
4. Preserve the reconstruction and simulation software and basic level data	Full potential of the experimental data

Table 1 The four levels of research data in HEP in order of increasing complexity<sup>70</sup>

### 2.1.2.1 Activity: HEPData integration in INSPIRE

INSPIRE<sup>71</sup> is the digital library and first point of information in the field of HEP. It is a co-operation between CERN<sup>72</sup>, DESY<sup>73</sup>, Fermilab<sup>74</sup> and SLAC<sup>75</sup>. INSPIRE provides not only literature but also additional tools to support researchers' daily work like citation analysis tools or a portal to job vacancies in the field.

Additionally to publications, it recently integrated additional data from the only data repository in HEP. These are tables that are displayed in a tab labelled with the source they are coming from which is complementary to references, citations and plots in context with the publication they belong to. As a value-added service, a Digital Object Identifier (DOI) will be assigned to most of the data sets on INSPIRE. This DOI assignment will be done in co-operation with the international initiative

<sup>68</sup> <http://www.dphep.org/>

<sup>69</sup> <http://hepdata.cedar.ac.uk/>

<sup>70</sup> DPHEP Study Group, 2009.

<sup>71</sup> <http://inspirehep.net>

<sup>72</sup> <http://www.cern.ch>

<sup>73</sup> <http://www.desy.de/>

<sup>74</sup> <http://www.fnal.gov/>

<sup>75</sup> <http://www.slac.stanford.edu/>

DataCite<sup>76</sup>. DOIs can be used to track the reuse of these datasets. In addition versioning of DOIs facilitates the identification of the most current data version. This will be done on INSPIRE as citations for data will be part of the overview of citation metrics provided.

Metadata quality is an important issue on the INSPIRE platform. The datasets are submitted with sufficient metadata to fulfil the mandatory requirements from DataCite to assign DOIs.<sup>77</sup> The data are submitted to the data repository HEPData, which is very well known in the community. There is no submission interface but submission is done via mail or tailored solutions for bigger files, so submitting data is actively done by data creators. Therefore, the repository staffs are always in contact with the submitters and other community members which eases getting high quality metadata.

The submitted data are checked for format compatibility by the HEPData staff. If the format does not match the community standard, they are transformed. The data repository provides export possibilities to commonly used visualisation or analysis platforms. These format variations are created on the fly and not stored and preserved additionally. Furthermore, the repository staff members provide summaries for some data sets as an additional service.

An important part of the data is the describing of the table headers. They provide information about the type of physical process as well as further description about the circumstances under which the experiment was carried out. These mathematical formulas were written by the HEPData staffs in a special text-based notation (Illustration 3) and are displayed better readable on INSPIRE (Illustration 4). In both cases however, they are fully searchable.

<b>PT : &lt; 15 GeV</b>
<b>YRAP : 2.0-4.5</b>
<b>SQRT(S) : 7000.0 GeV</b>
<b>SIG IN NB</b>

**Illustration 3 Header of a data table on the data repository site**

$p_T < 15 \text{ GeV}$
$y = 2.0 - 4.5$
$\sqrt{s} = 7000.0 \text{ GeV}$
$\sigma \text{ (NB)}$

**Illustration 4 Header of a data table on INSPIRE**

<sup>76</sup> <http://datacite.org>, see chapter 2.2.1

<sup>77</sup> These are the properties: identifier, creator, title, publisher, and publication year. Yet, DataCite allows more properties and as comprehensive metadata information as possible should be provided - in case it was submitted with the data. For example, it is possible to include versioning information and information about licences that are connected with the data sets. The property "version" is especially useful as there might be updated or corrected versions of data sets and all versions should be available and citable. Therefore, there will be different DOIs for the several versions of one dataset.

For the community, it is important that a trusted third party manages the data so that integrity is assured. The data repository has been an appreciated platform for decennia. INSPIRE became a trusted institution during the last years through providing different services to the community and researchers are very eager to see INSPIRE offering even more services and tools. INSPIRE already provides different possibilities for the community to improve the data provided, like updating references for papers.

In the future, INSPIRE will be connected to projects like ORCID<sup>78</sup> and its metadata will hence be complemented by links to these other services.

### 2.1.2.2 References and further reading:

DPHEP Study Group. (2009). Data Preservation in High-Energy Physics. Arxiv preprint arXiv:0912.0255, (November), 1-18. Retrieved from <http://arxiv.org/abs/0912.0255>

DPHEP Study Group. (2012). Status Report of the DPHEP Study Group: Towards a Global Effort for Sustainable Data Preservation in High Energy Physics. Arxiv preprint arXiv:1205.4667, (May), 1-93. Retrieved from <http://arxiv.org/abs/1205.4667>

South, D. (2011). Data Preservation in High Energy Physics. Arxiv preprint arXiv:1101.3186, (May), 1-18. Retrieved from <http://arxiv.org/abs/1101.3186>

### 2.1.3 Example: Research and development activities of FORTH-ICS

There is a trend towards semantic descriptions (e.g. semantic annotations over documents, or ontology-based annotations of various scientific data). These descriptions are expressed using elements from one or more metadata schemas or ontologies.

However, Semantic Web Ontologies are not static but evolve as the understanding of the domain (or the domain itself) grows or evolves. This evolution happens independently of the ontological instance descriptions (for short metadata) which are stored in the various Metadata Repositories (MRs) or Knowledge Bases (KBs). However, it is a common practice for a MR/KB to periodically update its ontologies to their latest versions (e.g. for reasons of interoperability). This is done by "migrating" the available metadata to the latest version of the ontology. (See Illustration 5 and Illustration 6)

---

<sup>78</sup> <http://about.orcid.org/>

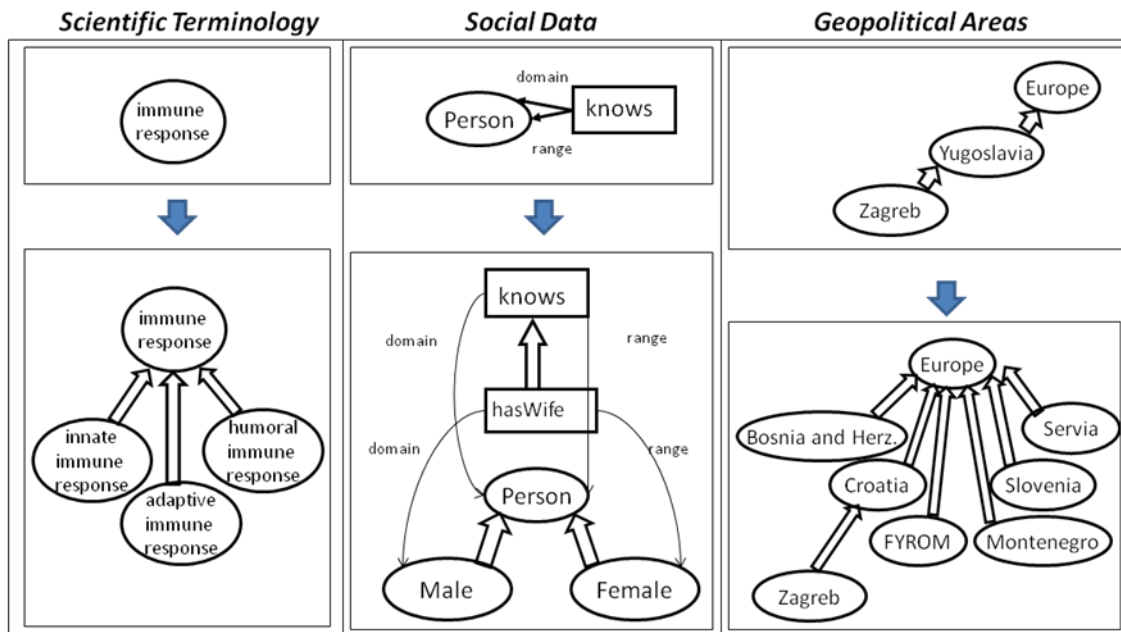


Illustration 5 Examples for schema/ontology evolution (a)

## Cultural Documentation

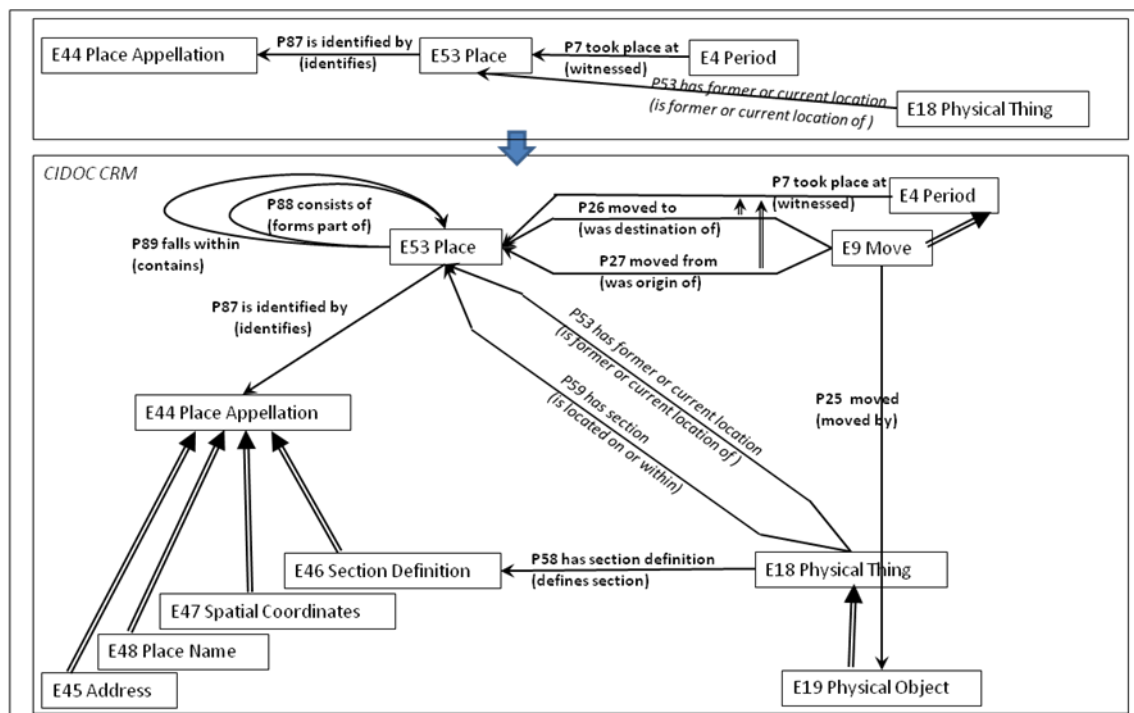


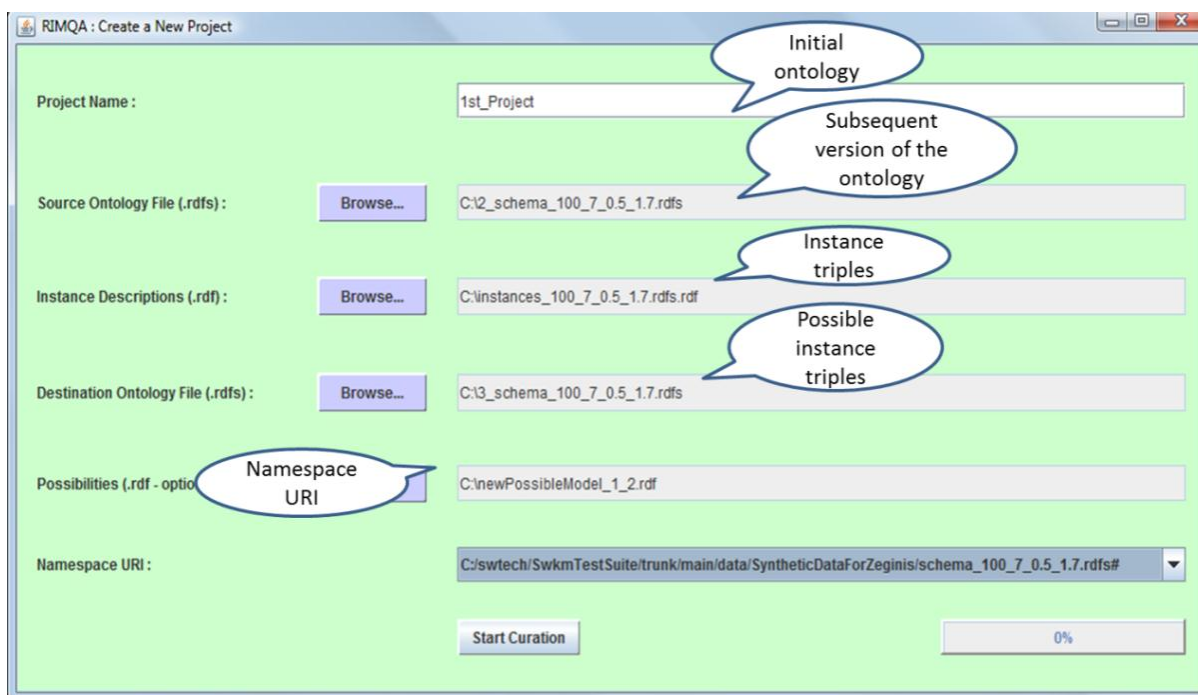
Illustration 6 Examples for schema/ontology evolution (b)

Usually such migrations are not difficult because new ontology versions are usually compatible with the past versions. However such migrations incur gaps regarding the specificity of migrated metadata. This results in inability to distinguish those metadata that should be re-examined for possible specialization (as consequence of the migration) from those for which this is not necessary. For this reason there is a need for principles, techniques and tools that can manage the uncertainty incurred by such migrations, specifically techniques which can identify automatically the descriptions that are candidate for specialization, compute, rank and recommend possible specializations, and flexible interactive techniques for updating the metadata repository (and its candidate specializations), after the user (curator) accepts/rejects such recommendations. This problem is especially important for "curated" KBs which have increased quality requirements.

FORTH aims at developing principles, techniques and software tools for tackling such issues.

A prototype tool (called RIMQA – RDF Instance Migration Quality Assistant) has already been implemented (some indicative screen dumps are given in illustrations X-Y) and the current research results were submitted and accepted for presentation and publication at *iPres2012*<sup>79</sup>:

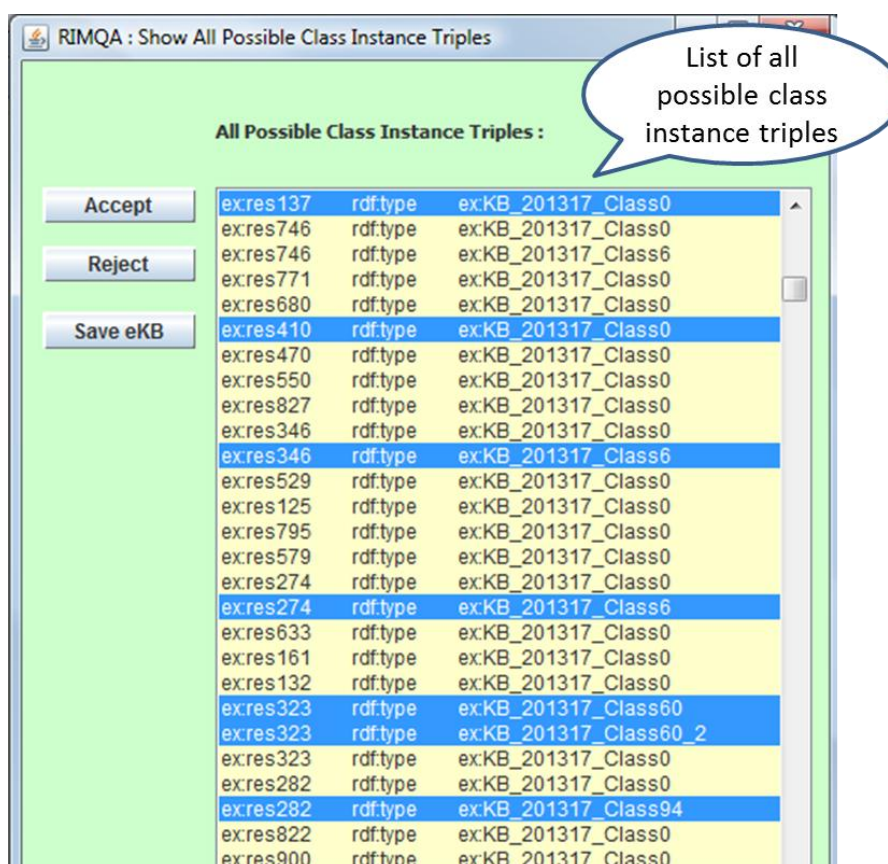
A prototype tool (called RIMQA – RDF Instance Migration Quality Assistant) has already been implemented and the current research results will be submitted for publication to *iPres2012* (See Illustration 7 - Illustration 9).



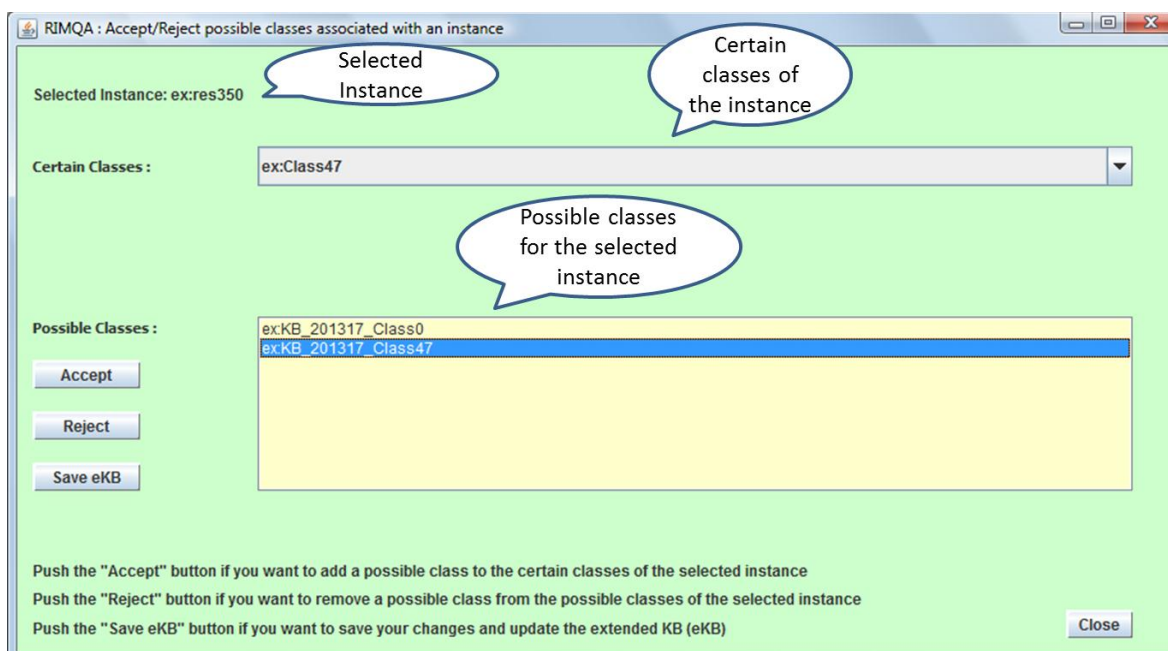
**Illustration 7 RIMQA - Start screen**

<sup>79</sup> Y. Tzitzikas, A. Analyti and M. Kampouraki, *Curating the Specificity of Metadata while World Models Evolve*, *Proceedings of the 9th Annual International Conference on Digital Preservation* (iPres2012), Oct. 2012, Toronto.





**Illustration 8 RIMQA - Recommendation of possible refinements**



**Illustration 9 RIMQA - Recommendation of possible refinements for one particular object**

## 2.2 Reputation

Reputation is a key currency in the science community. As of now there is no established system to reward scientists who make their research data available for reuse. One cornerstone to build such a system is an infrastructure that makes research data citable as is possible with publications.

The German project "Publication and Citation of Scientific Primary Data" was an early (2003-2005) approach tackling this challenge by developing identifiers which could be attached to data sets. Since 2009 DataCite made this approach operational, on a global scale.

The Data Archiving and Networked Services developed and operates "online archiving system" EASY which includes a module enabling recommending archived research data sets.

The two above mentioned projects provide "research journal functionalities" but stop short of actually starting a data journal. "Earth System Science Data Journal" is pilot project for data journals.

### 2.2.1 Example: STD-DOI and DataCite (Helmholtz Association)

Funded by the German Research Foundation DFG for the period 2003-2005, the project "Publication and Citation of Scientific Primary Data" (STD-DOI)<sup>80</sup> aimed to make research data citable as publications. The STD-DOI project used persistent identifiers to identify data sets. Persistent identifiers provide the opportunity to improve findability and accessibility of research data. One purpose of the project was to advance methods of data citation. "A citation of a data set adheres to the classical citation rules in scientific literature, e.g. author(s), publication year, data set name, persistent identifier."<sup>81</sup> To improve the citation of data sets a Digital Object Identifier (DOI), a persistent identifier, was used. The use of persistent identifiers is also considered to enhance the reputation of the data producers.

Since 2009 the international initiative DataCite<sup>82</sup> "helps researchers to find, access, and reuse data". The initiative builds on the experience of the STD DOI project. Extract from the DataCite project description:

"By working with data centres to assign persistent identifiers to datasets, we are developing an infrastructure that supports simple and effective methods of data citation, discovery, and access. Citable datasets become legitimate contributions to scholarly communication, paving the way for new metrics and publication models that recognise and reward data sharing."<sup>83</sup>

The following example shows the citation of a data set in an article<sup>84</sup> in the journal "Palaeogeography, Palaeoclimatology, Palaeoecology":

---

<sup>80</sup> <http://www.std-doi.de>

<sup>81</sup> Klump, J., Bertelmann, R., Brase, J., Diepenbroek, M., Grobe, H., Höck, H., Lautenschlager, M., et al. (2006). Data publication in the open access initiative. *Data Science Journal*, 5, 79-83. doi:10.2481/dsj.5.79

<sup>82</sup> <http://datacite.org>

<sup>83</sup> <http://datacite.org/whatdowedo>

<sup>84</sup> Bruch, A. A., Uhl, D., & Mosbrugger, V. (2007). Miocene climate in Europe — Patterns and evolution A first synthesis of NECLIME. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 253(1-2), 1-7. doi:10.1016/j.palaeo.2007.03.030



#### References of data available online in the PANGAEA data base

Akgün et al., 2007 Akgün, F., Kayseri, M.S., Akkiraz, M.S., 2007. Neogene palaeoclimate reconstructions in Anatolia (Turkey). PANGAEA. doi:10.1594/PANGAEA.596351.. | View Record in Scopus | Cited By in Scopus (26)

Böhme et al., 2007 Böhme, M., Bruch, A.A. Selmeier, A., 2007. Miocene palaeoclimate reconstructions from the North Alpine Foreland Basin in Germany.

### Illustration 10 Example of a data citation

The cited data set is accessible on the PANGAEA<sup>85</sup> data repository:

"Each dataset can be identified, shared, published and cited by using a Digital Object Identifier (DOI). Data are archived as supplements to publications or as citable data collections. Citations are available through the portal of the German National Library of Science and Technology (GetInfo)."<sup>86</sup>

Such procedures are also supported by other data repositories like DRYAD or ICDP Scientific Drilling Database:

- DRYAD: "Dryad is an international repository of data underlying peer-reviewed articles in the basic and applied biosciences. Dryad enables scientists to validate published findings, explore new analysis methodologies, repurpose data for research questions unanticipated by the original authors, and perform synthetic studies."<sup>87</sup>
- ICDP Scientific Drilling Database: "The Scientific Drilling Database is operated by the ICDP Operational Support Group and the Data Center of GeoForschungsZentrum Potsdam. It holds data resulting from ICDP projects and other projects supported by the ICDP Operational Support Group."<sup>88</sup>

### 2.2.2 Example: ODE project (several APARSEN partners)

Some APARSEN partners explore the practices of data citation in the ODE project.

"The project will identify, collate, interpret and deliver evidence of emerging best practices in sharing, re-using, preserving and citing data, the drivers for these changes and barriers impeding progress, in forms suited to each audience."<sup>89</sup>

In 2011 the ODE project released a comprehensive report on "Integration of Data and Publications". The report describes credit as an important incentive for data sharing: "Researchers need to get credit for data as a first class research object"<sup>90</sup>. This finding is widespread across the disciplines. The editors of Nature Biotechnology already brought it to the point in 2009:

<sup>85</sup> <http://www.pangaea.de>

<sup>86</sup> <http://www.pangaea.de/about/>

<sup>87</sup> <http://datadryad.org/about>

<sup>88</sup> [http://www.scientificdrilling.org/front\\_content.php?idcat=239](http://www.scientificdrilling.org/front_content.php?idcat=239)

<sup>89</sup> <http://ode-project.eu>

<sup>90</sup> Reilly, S., Schallier, W., Schrimpf, S., Smit, E., & Wilkinson, M. (2011). Report on Integration of Data and Publications. Retrieved from <http://ode-project.eu/outputs>

"Data DOIs would not only enhance a researcher's reputation but also establish priority of data generation. Most important of all, they would provide a way to acknowledge the time and effort individuals must invest in sharing data, which ultimately benefits the scientific community as a whole."<sup>91</sup>

## 2.2.3 Example: EASY (DANS)

The APARSEN partner, Data Archiving and Networked Services (DANS), has enabled commenting of datasets stored in the "online archiving system" EASY, in accordance with pre-defined criteria, since 2010. EASY enables access "to thousands of datasets in the humanities, the social sciences and other disciplines. EASY can also be used for the online depositing of research data."<sup>92</sup> The assessment of a dataset becomes visible for the user, if two assessments have been submitted for a dataset. A ranking system shows the reputation of a dataset.<sup>93</sup>

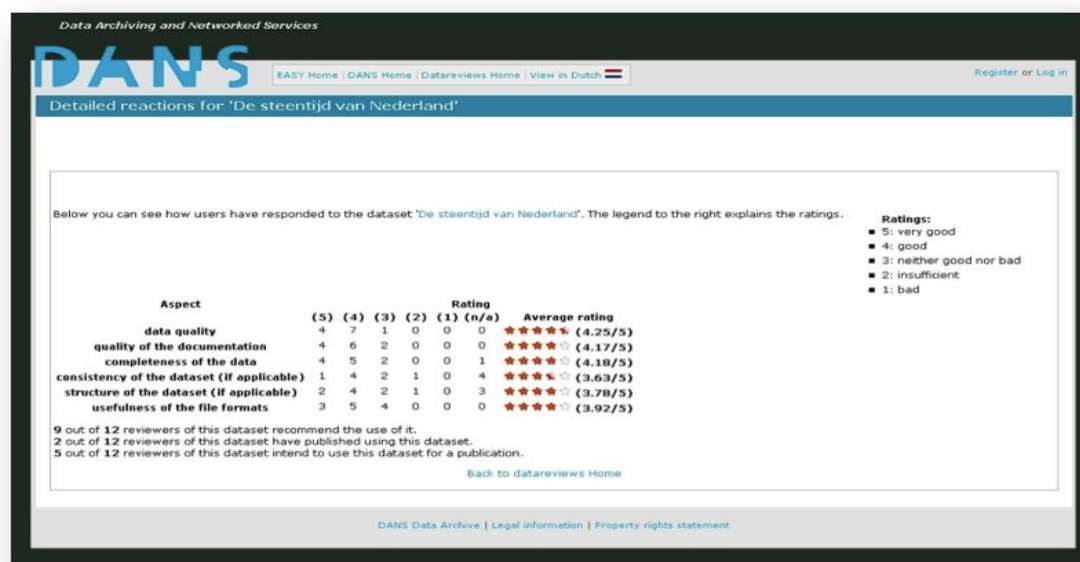


Illustration 11 Assessment of the dataset "De steentijd van Nederland"

## 2.3 Data quality

Data quality is a crucial factor when building an infrastructure for long term data preservation and reuse. The scientists' interest in depositing their research data in such an infrastructure and to reuse data being made available via one is closely linked to their persuasion that acceptance of data into the

<sup>91</sup> Credit where credit is overdue. (2009). Nature biotechnology, 27(7), 579. doi:10.1038/nbt0709-579

<sup>92</sup> <https://easy.dans.knaw.nl>

<sup>93</sup> A detailed description can be found at: Data Archiving and Networked Services. (2011). Data Reviews. Peer-reviewed research data. Retrieved from <http://www.dans.knaw.nl/en/content/categorieen/publicaties/dans-studies-digital-archiving-5>

system is an indication of quality and that the reputation of findings based on the reuse of data is not compromised by doubt concerning the quality of the data source.

In the context of long term data preservation the issue of quality is not limited to the data which are to be preserved for reuse but relate equally to the infrastructure which is to preserve these data and make them available for reuse.

This double challenge is reflected in the multi-faceted program "Quality Assurance Framework for Earth Observation" currently being developed and implemented by the European Space Agency. A national project, MaNIDA, faces the challenge and chance to reengineer high quality data management from sensor to dissemination on a large scale, in a complex social environment. A journal, ESSD, ties to establish quality checking on those data not yet emerging from such systematic environments.

### **2.3.1 Example: MaNIDA - Coordinated provision of quality information (Helmholtz-Association)**

In Germany, a number of funders and operating institutions are responsible for a sizable fleet of large and mid-sized research ships<sup>94</sup>, representing a huge investment and many millions of Euro in operating cost per year. The following discussion should be viewed considering that three of the project partners are Helmholtz centers, operating large research infrastructures (ships, airplanes, fixed monitoring systems underwater and at sea, and a long term data repository), two universities and one government agency (with another long term preservation mission).

In late 2011, a Helmholtz project, MaNIDA, was kicked off which is to provide two major outcomes. The most visible one will be to provide researchers with access to all data from German marine research, particularly from the ships mentioned above – at the insistence of researchers and funders alike. More behind the scenes, there is also the major task to define and implement common quality and quality assurance methods.

It is actually the easy part to provide the technology to harvest metadata and to make them searchable and browsable. This became clear even to those who had not previously been involved in long term preservation of data, after a brief period.

The primary and very hard problem is to agree on and implement a common naming of the parameters measured – possibly according to (emerging) international standards - and to harmonize conditions of access. In the context of access it was considered for a brief period to store digitized diplomatic letters along with data, which for each expedition and each country affected lay out which parameters may be measured in its exclusive economic zone and perhaps impose restrictions on the use of such data.

There is a huge variety of other attributes – like naming and numbering of expeditions – which need to be harmonized in order to implement useful search “facets”<sup>95</sup> and to detect (potential) duplicates.

Regarding quality, even before describing processes to assure a common or at least similar quality of results, the institutions involved need to find common ground on the “descriptors for data quality and processing levels”. Among those are, e.g. “quality flags”. In illustration 7 one of the contributing projects, COSYNA, documents its use of an attribute indicating quality assurance. Note that one of the problems addressed is the need to deviate from an existing international convention.

---

<sup>94</sup> <https://www.portal-forschungsschiffe.de/schiffe>

<sup>95</sup> [http://en.wikipedia.org/wiki/Faceted\\_search](http://en.wikipedia.org/wiki/Faceted_search)

### 3.1.1 Quality Flag Scheme

COSYNA uses a uniform scheme of Quality Flags (QF) applicable to generated/processed data for physical, chemical and biological parameters:

QF	Definition	Criteria
0	No quality control	no QC applied
1	Good data	delayed QC passed
2	Probably good data	all NRT QC checks passed
3	Probably bad data	NRT QC not passed, data potentially correctable
4	Bad data	NRT and/or delayed QC not passed
9	Missing data	data not available


*Remark:* The coding (QF) and the definitions are a subset of the SeaDataNet QF scheme<sup>1</sup> and comply with the recommendations of EUROGOOS/DATA-MEQ<sup>2</sup> and MyOcean<sup>3</sup> for Near-Real Time Quality Control (NRT-QC). The criteria have been chosen in order to establish one uniform and unambiguous scheme applicable to the complete process of NRT and delayed QC. Thus, the criteria differ partly from the above mentioned recommendations, e.g. QF 1 (good data) is only set after delayed QC has been passed successfully.

### Illustration 12 Quality Flag Scheme of the COSYNA project<sup>96</sup>

Similarly, there is discussion even about a coarse naming for processing levels: One might think that something like “raw data”, “(calibrated, quality controlled) primary data” and “derived data” might be good enough. But here a distinction between what can and needs to be done in near real time and what should be done before archiving is needed – a distinction completely lost on the majority of MaNIDA participants not previously involved in NRT-processing and -dissemination of data.

Regarding data levels, COSYNA – and perhaps MaNIDA, following their lead – is oriented at the definitions from remote sensing (ESA) – but sees the need to differentiate here as well, due to the NRT challenge, see Illustration 13 below.

<sup>96</sup> “COSYNA Quality Assurance Framework”, (2011), unpublished technical document of COSYNA - Coastal Observation System for Northern and Arctic Seas, [http://www.hzg.de/institute/coastal\\_research/cosyna/](http://www.hzg.de/institute/coastal_research/cosyna/)

	<p style="text-align: center;"><b>COSYNA</b> <b>Datamanagement</b></p>	<p>Doc. No : CO-DM-001 Issue: 1 Rev: 0.14 Name : Common level structure for data used within COSYNA Date : 2012-03-09 Page : 2</p>
---	--	--

DL	Definition	Typical User / Access	Comment
0	Raw data (cps, a.u.)	Data Generator & Processor	Storage for QA purposes and reprocessing
1	Parameters (phys., chem. biol.) (ASCII or data-specific)	Scientists other than Data Originator	Access permitted after request from Data Originator
2	Parameters space-time-referenced (ASCII, NetCDF or data-specific)	Scientists other than Data Originator	Access permitted after request from Data Originator
3	Data Products NetCDF (maplike), RDBMS (timeseries)	World / Access via COSYNA-Portal	Access open under Cosyna policies – data are subject to changes (corrections)
4	Published Data Products DOI und Storage e.g. in Pangea	World / Access e.g. via Pangea	Citation via DOI, long-term storage, data not changable

### Illustration 13 Data Level - Overview of the COSYNA project<sup>97</sup>

Only after agreeing on these concepts and their encoding it will become possible to talk about the actual “QC checks” to be applied for each parameter – and possibly to provide proof of who applied them when. In the end, after much hesitation over how to translate information available in each repository, it will be of utmost interest to see that many datasets held are actually raw – the QC-ed data being withheld by researchers – or that two repositories hold a dataset at different QC or processing levels, which could have been confused with being duplicates, previously.

It is worthwhile noting that much of the work needed is to overcome hesitation and “Angst” – which may be well founded in a) lack of resources to push through all the necessary changes and additions and b) the apprehension that something might go wrong due to operating on a shoestring and cause irreparable damage to the valuable data holdings, built over decades.

The project profits immensely from the fact that nine ships use almost identical versions of one data acquisition and management system, “DShip”<sup>98</sup>, and data end up in just two long term data repositories, PANGAEA and BSH/DOD, and one NRT data distribution system<sup>99</sup>. But the nine ships are managed and operated by a large number of stakeholders<sup>100</sup> and an even larger number of expedition leaders and groups’ principal investigators who will need to get the message and adhere to processes and conventions, once they are adopted. In the probably best of outcomes, data specialists from repositories, NRT data lab and portals will have to continue their current work of coordination of

<sup>97</sup> “Common level structure for data used within COSYNA”, (2012), unpublished technical document of COSYNA - Coastal Observation System for Northern and Arctic Seas, [http://www.hzg.de/institute/coastal\\_research/cosyna/](http://www.hzg.de/institute/coastal_research/cosyna/)

<sup>98</sup> “Werum equips new “Sonne” Research Vessel with Data Management System”  
[http://www.werum.de/en/mdmnews/news/NR\\_TFS-Sonne-Nachf\\_2012-07-31.jsp](http://www.werum.de/en/mdmnews/news/NR_TFS-Sonne-Nachf_2012-07-31.jsp)

<sup>99</sup> [www.pangaea.de](http://www.pangaea.de), [www.bsh.de/en/Marine\\_data/Observations/DOD\\_Data\\_Centre/](http://www.bsh.de/en/Marine_data/Observations/DOD_Data_Centre/) and [coastlab.org](http://coastlab.org)

<sup>100</sup> as explained for RV Polarstern in the context of proposals for ship-time:  
[www.awi.de/en/infrastructure/ships/polarstern/submission\\_of\\_proposals/](http://www.awi.de/en/infrastructure/ships/polarstern/submission_of_proposals/)

requirements as a perhaps even permanent task of advising (if not educating) the same group: scientists, ships' crews as well as policy setting bodies about the outcomes.

### 2.3.2 Example: Earth System Science Data Journal (Helmholtz-Association)

This description is based on excerpts from the journal site<sup>101</sup> and an article by the chief editors.<sup>102</sup>

Earth System Science Data (ESSD) is an international, interdisciplinary open access journal for the publication of articles on original research data sets. The editorial board encourages submissions of original data or data collections which are of sufficient quality and potential impact to contribute to these aims.

ESSD has an innovative two-stage publication process involving the scientific discussion forum Earth System Science Data Discussions (ESSDD), which has been designed to:

- foster scientific discussion;
- maximise the effectiveness and transparency of scientific quality assurance;
- enable rapid publication of new scientific results;
- make scientific publications freely accessible.

In the first stage, papers that pass a quick editorial review are immediately published as discussion papers on the ESSDD website. They are then subject to "interactive public discussion", during which the referees' comments (anonymously or attributed), additional short comments by other members of the scientific community (attributed) and the authors' replies are also published openly in ESSDD. In the second stage the final revised papers are published in ESSD, if the peer-review process is completed and the paper accepted.

The precondition to submit a manuscript for publication in ESSD and its scientific discussion forum ESSDD is that the data sets referenced in the manuscript are submitted to a long-term data repository.

For future reuse and reinterpretation it is mandatory for the user to be assured about research data quality. It is the aim of ESSD to provide the quality assessment for datasets which already reside in permanent repositories.

The data must be presented readily and accessible to inspection and analysis to make the reviewer's task possible. Even if a dataset submitted is the first ever published (on a parameter, in a region, etc.), its claimed accuracy, the instrumentation employed and methods of processing must reflect the "state of the art" or "best practises". Considering all conditions and influences presented in the article, these claims and factors must be mutually consistent. The reviewers will then apply their expert knowledge and operational experience in the specific field and may perform tests, e.g., statistical tests, to make a judgement whether the claimed findings and its factors – individually and as a whole – are plausible and without detectable faults.

To ensure publication precedence for authors, and to provide a lasting record of scientific discussion, ESSDD and ESSD are both ISSN-registered, permanently archived and fully citable.

---

<sup>101</sup> <http://www.earth-system-science-data.net>

<sup>102</sup> Pfeiffenberger, H., & Carlson, D. (2011). "Earth System Science Data" (ESSD) - A Peer Reviewed Journal for Publication of Data. D-Lib Magazine, 17(1/2). doi:10.1045/january2011-pfeiffenberger



### 2.3.3 Example: Remote sensing domain (ESA)

ESA introduces in this section an insight on the relation existing between the quality of data and their preservation, which will be the key topic for the following text.

The domain is that of Earth Observation (EO) and the activity ESA is currently carrying on this topic is called "Evaluation of requirements on data quality information in relation the Long Term Data Preservation (LTDP) guidelines", funded by ESA's General Studies Programme<sup>103</sup>. Recent developments in ESA's Earth Observation Programmes in the field of data quality assurance and long term data preservation are the background to this activity.

#### 2.3.3.1 Context

In recent years the need for accessing historical Earth Observation (EO) data series strongly increased, mainly for long term science and environmental monitoring applications. This trend is likely to increase even more in the future in particular due to the growing interest on global change monitoring that requires data time-series spanning 20 years and more, and for the need to support the United Nations Framework Convention on Climate Change (UNFCCC).

Content of EO space data archives is extending from a few years to decades and their scientific value is continuously increasing hence is well recognized the need to preserve them without time limitation and to keep the archived EO space data well accessible and exploitable as they constitute a humankind asset. In addition, the large amount of new Earth Observation missions upcoming in the next years will lead to a major increase of EO space data volumes. This fact, together with the increased demands from the scientific user community, marks a challenge for Earth Observation satellite operators, Space Agencies and EO space data providers regarding coherent data preservation and optimum availability and accessibility of the different data products. Through a cooperative and harmonized collective approach at European level coordinated by ESA with the involvement of the Ground Segment Coordination Body (GSCB), these needs led to the establishment of the Earth Observation Long Term Data Preservation (LTDP) set of guidelines<sup>104</sup>.

Preservation requires the availability of common approaches based on established international procedures and policies aimed at guaranteeing that data are kept appropriately and that they will still be available in the future. In addition to the data, also the associated knowledge should be duly preserved to guarantee their comprehension and usability in future. The Earth Science domain is a clear example where such common procedures and policies are missing and where the future accessibility and usability of data is at risk. It is important to point out that the "knowledge" associated to Earth Science data is not yet fully consolidated. A first attempt to define the additional information to be preserved in the long term in addition to the primary data to allow the future exploitability of the data has been done in the Earth Observation domain and resulted in a "Content Standard document" that needs further consolidation.

The picture is made more complicated by the fact that it is also not known at the moment how data may be used by future scientists and researchers and therefore the preservation of information that seems not useful today might be of great importance for future generations of users.

At the same time, the Group on Earth Observations (GEO) has recognised the necessity to develop a data quality assurance strategy that guarantees the correct applicability and optimises the

---

<sup>103</sup> [www.esa.int/gsp](http://www.esa.int/gsp)

<sup>104</sup> <http://earth.esa.int/gscb/ltdp/index.html>

interoperability of EO products acquired by a large variety of sources across missions and sensors whether space-borne or on the ground. In response to this need, the Committee on Earth Observation Satellites (CEOS) Working Group on Calibration and Validation (WGCV) established and endorsed the Quality Assurance Framework for Earth Observation<sup>105</sup> guidelines.

The QA4EO is composed of a set of 7 guidelines based on a core principle: "a measurement/process must have associated with it a Quality Indicator (QI) based on documented metrological assessment of its traceability to community agreed absolute reference standards". The objective of QA4EO is to assure the correct applicability and interoperability of EO products.

For EO missions, the metrological traceability to absolute calibrated reference standards of the delivered product is a complex objective which requires tackling the phenomena to be measured, the platform, the instrument, and all auxiliary applied models and processing algorithms.

In the above framework, recent workshops<sup>106</sup> have highlighted the strong ties between the LTDP and QA4EO guidelines, showing a need to systematically evaluate the preservation requirements generated with respect to the data quality information.

### 2.3.3.2 Activity

It is worth noting that at the time of writing the activity is started but still at the very beginning, and therefore results might be available well after this deliverable submission to EC.

Information referred as "Secondary data" by the EO European LTDP Common Guidelines needs to be preserved together with the so called "Primary data", in order to allow present and future understanding and usability of the data. Among this information, data quality is a fundamental one.

Ultimately, the aim of this ESA's funded activity is to preserve all this information with particular attention to quality information and specifications and all related auxiliary data that will allow the correct applicability of the payload data without time limitation. This means maintaining the full history of the dataset and the capability to re-process data, therefore the understanding of the principles of the measurements from the instrument to the processor. The "Secondary data" defined in the Earth Observation European LTDP Common Guidelines need to include all the data quality related information.

One of main objectives of this activity is to assess case studies applying the LTDP and QA4EO guidelines and Content Standard document to EO and to identify specific needs, critical aspects and potential improvements.

Going into details, for the EO field, the current situation for the following mission / instruments will be investigated during this activity:

- ERS 1 / Radar Altimeter (RA)
- ERS 2 / RA
- Envisat / RA-2
- ERS 1 / Along Track Scanning Radiometer (ATSR)
- ERS 2 / ATSR
- Envisat / Advanced ATSR (AATSR)
- ERS 1 / Synthetic Aperture Radar (SAR)

---

<sup>105</sup> QA4EO - <http://qa4eo.org/>

<sup>106</sup> e.g. <http://earth.esa.int/gscb/lt dp/objectives-workshop2010.html>



- ERS 2 / SAR
- Envisat / Advanced SAR (ASAR)
- Envisat / Global Ozone Monitoring by Occultation of Stars (GOMOS)
- Envisat / Michelson Interferometer for Passive Atmospheric Sounding (MIPAS)
- Sentinel-2 / Multi-Spectral Imager (MSI)

Part of the work consists in applying the LTDP guidelines and Content Standard document tailoring them to the specificities of the mission/instruments and identifying the respective relevant "Secondary data" including all the quality information.

Implementation issues will be tackled as well, on the preservation of the "Secondary data", including the quality information. General implementation strategies that can be applied for past, current or future missions will be derived from this activity.

Finally LTDP and QA4EO guidelines and Content Standard document will be reviewed once more in order to transfer the knowledge acquired back into the guidelines themselves.

### 2.3.3.3 Useful references

Albani, M., Beruti, V., Duplaa, M., Giguere, C., Velarde, C., Mikusch, E., Serra, M., et al. (2010). Long Term Data Preservation of Earth Observation Space Data. European LTDP Common Guidelines. Issue 1.1. (M. Albani, Ed.). Retrieved from [http://earth.esa.int/gscb/ltdp/EuropeanLTDPCommonGuidelines\\_Issue1.1.pdf](http://earth.esa.int/gscb/ltdp/EuropeanLTDPCommonGuidelines_Issue1.1.pdf)

QA4EO Task Team. (2010). A Quality Assurance Framework for Earth Observation. Principles. Version 4.0. Retrieved from [http://qa4eo.org/docs/QA4EO\\_Principles\\_v4.0.pdf](http://qa4eo.org/docs/QA4EO_Principles_v4.0.pdf)

Long Term Preservation of Earth Observation Space Data: European LTDP Common Guidelines (<http://earth.esa.int/gscb/ltdp/index.html>)

QA4EO Guidelines (seven documents) (<http://qa4eo.org>)

### 3 Results of an internal survey

The APARSEN partners decided to conduct an internal survey in order to identify strategies for further research on the area of annotation, reputation and data quality.

The aim of this survey was to investigate the current practices and views of the APARSEN partners on these complex topics. The core question of this survey was:

How do the APARSEN partners deal within their organizations and especially their repositories with annotation, reputation and data quality?

The results also serve as a basis for further dialogue on topics like preservation services, cost models and business cases in the network of excellent.

The questionnaire was addressed to all member institutions of the APARSEN consortium. Answers were given from the perspective of a data repository.

A data repository was defined as a virtual facility for the deposit of electronic copies of scientific data. (In this survey we didn't focus on institutional open access repositories for the deposit of academic text publications, such as academic journal articles.)

The questionnaire was divided into four sections:

1. General questions about the data repository
2. Questions about annotation services
3. Questions about reputation
4. Questions about data quality

The online survey was open for six weeks during September and October 2011. The survey was conducted by Survey Monkey, an online survey tool.<sup>107</sup> The survey consisted of 54 questions (see Annex 10). Some of the questions were inherited from the PARSE.insight Survey<sup>108</sup> and the ELIXIR database provider survey<sup>109</sup>. 20 partners provided answers to the questionnaire. Taking into account that not every one of the 31 APARSEN members operates a data repository this is an agreeable response rate. The responses were aggregated and anonymised as assured to the respondents. Some of the answers were compared with respective results of the PARSE.insight survey.

The results give a comprehensive summary of the views of APARSEN partners on annotation, reputation and data quality.

In the following the results are analysed by four categories:

1. the APARSEN data repository landscape,
2. the field of annotation services,
3. the topic of reputation and
4. the subject of data quality.

The description of the APARSEN data repositories landscape should be also seen as a foundation for the further development of the network of excellence.

---

<sup>107</sup> <http://surveymonkey.com>

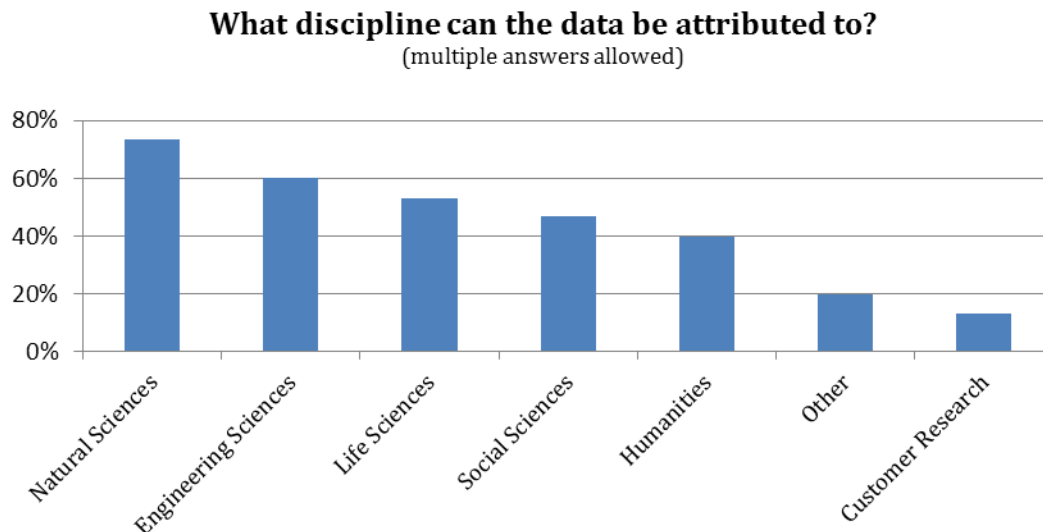
<sup>108</sup> Kuipers, T., & Van der Hoeven, J. (2009). Insight into digital preservation of research output in Europe. Survey Report. Framework. Retrieved from [http://www.parse-insight.eu/downloads/PARSE-Insight\\_D3-4\\_SurveyReport\\_final\\_hq.pdf](http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf)

<sup>109</sup> Southan, C., & Cameron, G. (2009). Database Provider Survey. Report for ELXIR Work Package 2 (30th ed.). Retrieved from [http://www.elixir-europe.org/bcms/elixir/Documents/reports/WP2\\_Annex-Provider\\_Survey\\_Report.pdf](http://www.elixir-europe.org/bcms/elixir/Documents/reports/WP2_Annex-Provider_Survey_Report.pdf)

### 3.1 Data repositories landscape

The questions in this section are the fundament of the survey. The answers describe the current landscape of data repositories in the APARSEN consortium.

In the first set of questions the disciplines and data types were identified. Multiple answers were permitted:



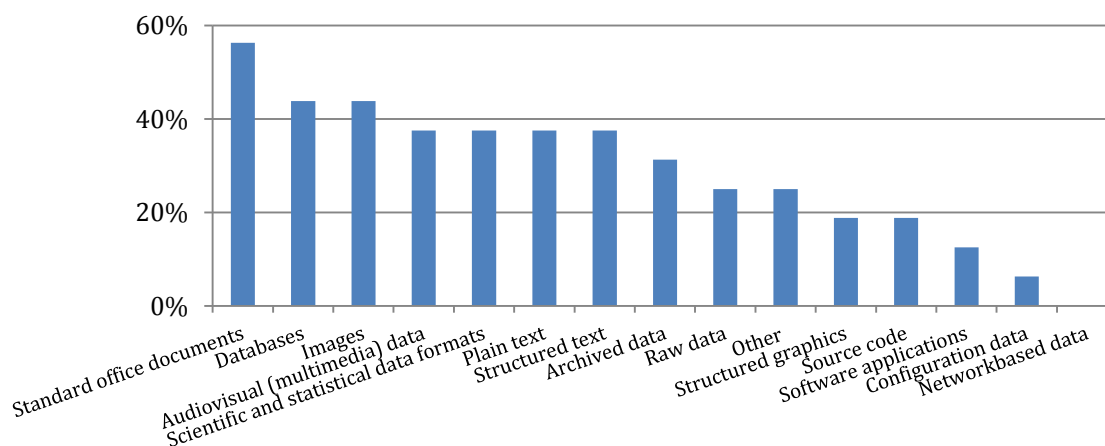
**Figure 1 Disciplines of stored data, n = 15**

Figure 1 shows that the majority of the repositories in the APARSEN Network can be assigned to the life, natural and engineering science (187%). Another focus is on the humanities and social science (87%).

Further, the respondents were asked about different kinds of data types that are stored in their data repositories (see Figure 2). Categories from the PARSE.insight survey were used for this. Standard office documents (text documents, spread sheets, presentations) were most frequently mentioned (56%). Interesting is the frequent mention of the category database (DBASE, MS Access, Oracle, MySQL, etc.) and images (JPEG, JPEG2000, GIF, TIF, PNG, SVG, etc.) (in sum 44%). The storage of data bases poses the challenge to enable their future use.

### Please indicate what type of data is stored in the repository.

(multiple answers allowed)



**Figure 2 Data types, n = 16**

In a further question the respondents were asked to relate the stored data to four categories: "research data", "governmental data", "cultural data", "internal company data" and "other". Multiple answers were permitted. About 87% of the respondents classified their data as "research data", followed by "cultural data" with 50%. This result reflects the two main groups of data producers in the APARSEN network: cultural institutions, like libraries and scholarly institutions, like research labs.

### 3.1.1 Data repository systems

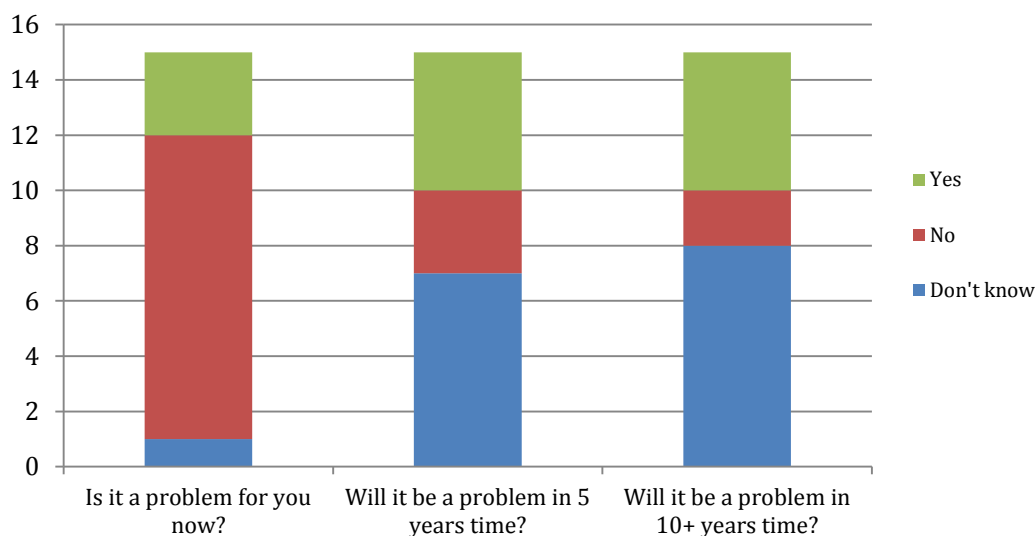
The survey also looked on the specifics of the APARSEN data repository landscape.

A set of questions dealt with the funding structure of the repositories. Two questions of the ELIXIR database provider survey were used for this purpose.

The answers to the question "What type of funding does your repository have?" (Multiple answers were permitted.) show the most common funding source is the institution, which runs the data repository (80%). Other funding sources are government grants: national (40%) and European (33%).

Interesting is the assessment of the future funding status of the repositories. Figure 3 shows that the future funding opportunities of the repositories are very uncertain. The PARSE.insight survey already indicated the high prevalence of short term projects in the field of data preservation and the ELIXIR survey found, that only 3.5% of the bimolecular databases are financed for longer than 5 years. These results demonstrate the need for long-term financing concepts to ensure the permanent access to data. Thus, the development of sustainable financing strategies is one of the major challenges.

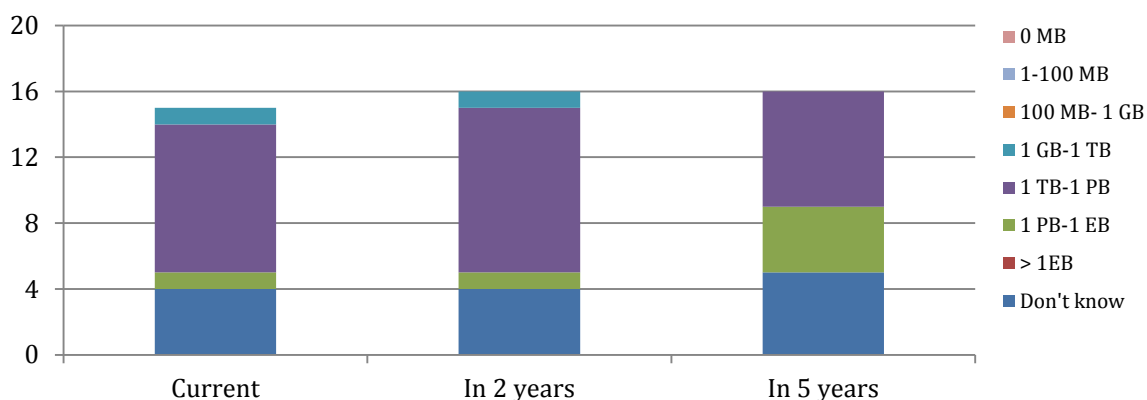
### Funding - now and in the future.



**Figure 3 Funding of the repositories - now and in the future, n = 15**

The discussion of the "data deluge" was taken up with a question about the amount of data stored. Currently most of the APARSEN data repositories work in the terabyte area. The respondents expect a growing amount of data in the next years (see Figure 4).

### What amount of data is stored in the repository? Please estimate also the volume in 2 and 5 years.



**Figure 4 Amount of stored data, n = 16**

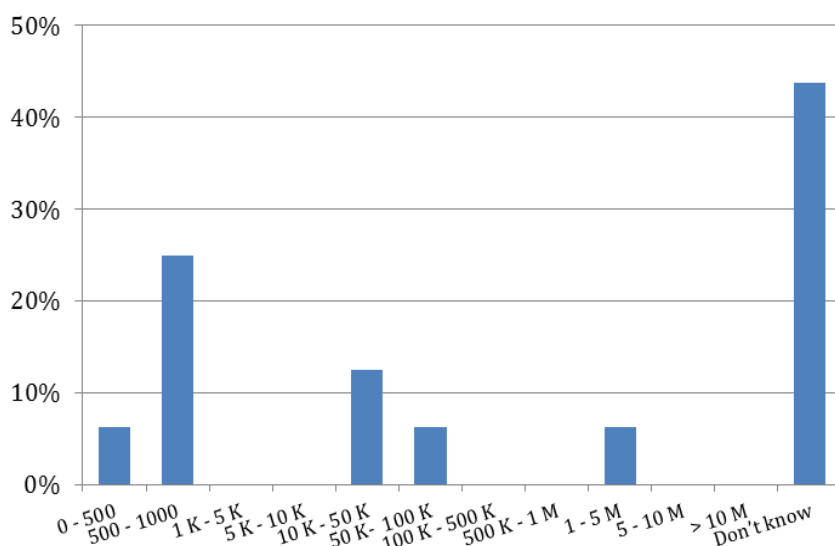
The required storage size is neither an authoritative indicator concerning the value of the stored data nor does it indicate the amount of work that was necessary to build up and to maintain the data

repository. This is perhaps the reason for the puzzling fact that 25% of respondents "don't know" the amount of stored data. Additional indicators need to be developed in order to appraise expected/legitimate costs of a data repository.

Preservation is the core topic for the APARSEN members. The survey asked about the most common preservation strategies. Normalization and Migration are the most frequently mentioned preservation strategies (both 54%), followed by emulation procedures (23%).

Further, the use of the data repository was discussed in the survey. The largest group of respondents is does not know the number of hits per month (44%). This result points to a lack of usage standards and monitoring tools. The largest group of the respondents who know the number of hits on their data repository have reported a usage of 500 - 1000 hits per month (25%). This indicates a very specialized group of data repository users.

**Where you can, please supply web hits per-month (excluding web-crawling).**



**Figure 5 Web hits per month, n = 16**

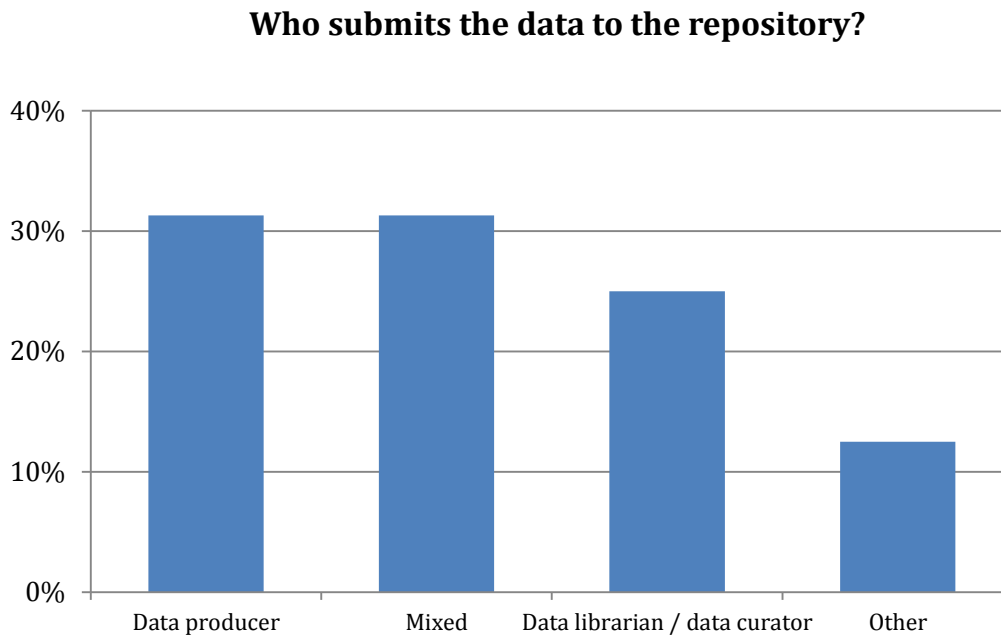
The development of usage measurement is indispensable to legitimize funding of data repositories. The development of standards to record visits and downloads is an important step in this direction. However, further or other indicators may be of greater importance. Obviously there is the value associated with stored data. Also usage volume needs to be related to the size of the pertinent research community of the collections stored in a data repository.

In some cases special knowledge is needed, to enable the re-use and re-purposing of stored data. Most of the APARSEN partners offer training, to ensure good practice when dealing with the stored data (75%).



### 3.1.2 Ingest and accessibility

The ingest process is a central step in the data management process. Figure 6 shows the distribution of the actors who are submitting data to the repository. The result shows the balanced relationship between data producers (31%) and data librarians/curators (25%). In addition automatic ingest processes were mentioned.



**Figure 6 Actors in ingest process, n = 16**

An important question concerns the integration of repositories in an institution or a specific focus group. The respondents were asked if the storage of data in their data repository is voluntary or mandatory. 37% of the respondents indicated that the storage is voluntary. Only 19% said that the storage is mandatory. The following two reasons were given for this mandatory practice:

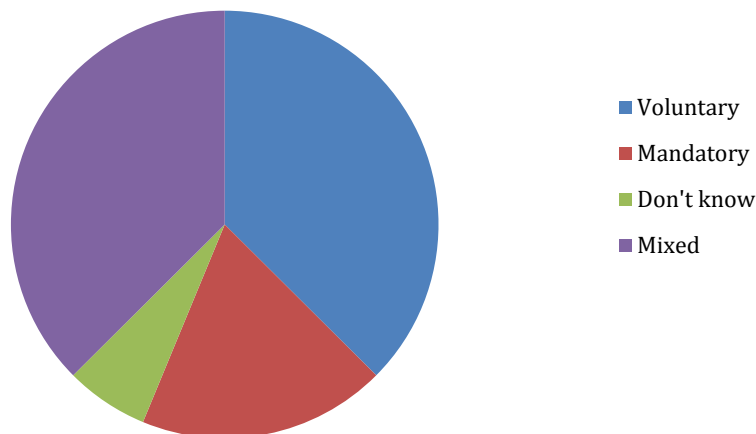
- Institutional mandates for the deposit of specific data types
- Data deposit mandates by funding bodies

Another 37% indicated different circumstances.

Where possible, data deposit should be prompted by incentives rather than mandates. In the APARSEN network, the following incentives to support the storage of data in a repository were mentioned:

- Better visibility of the data
- Funding of the data collection
- Long-term preservation services
- Trusting environment for the data

### Is the storage in the repository for the target group voluntary or mandatory?



**Figure 7 Storage - voluntary or mandatory, n=16**

Accessibility of data was the subject of a further question. In most cases the accessibility of the data depends on conditions attached to the stored data sets. Only 12% of the repositories allow free access to their whole content. 56% of the respondents indicated varying levels of data access. The following explanatory notes were given for restricted accessibility of the data:

- Only accessible after a three years embargo period
- Only accessible for a specific user group
- Only accessible for a data librarian or a data curator
- Only accessible in the institution which operates the data repository

Accessibility does not necessarily imply that downloading the data sets is permitted.

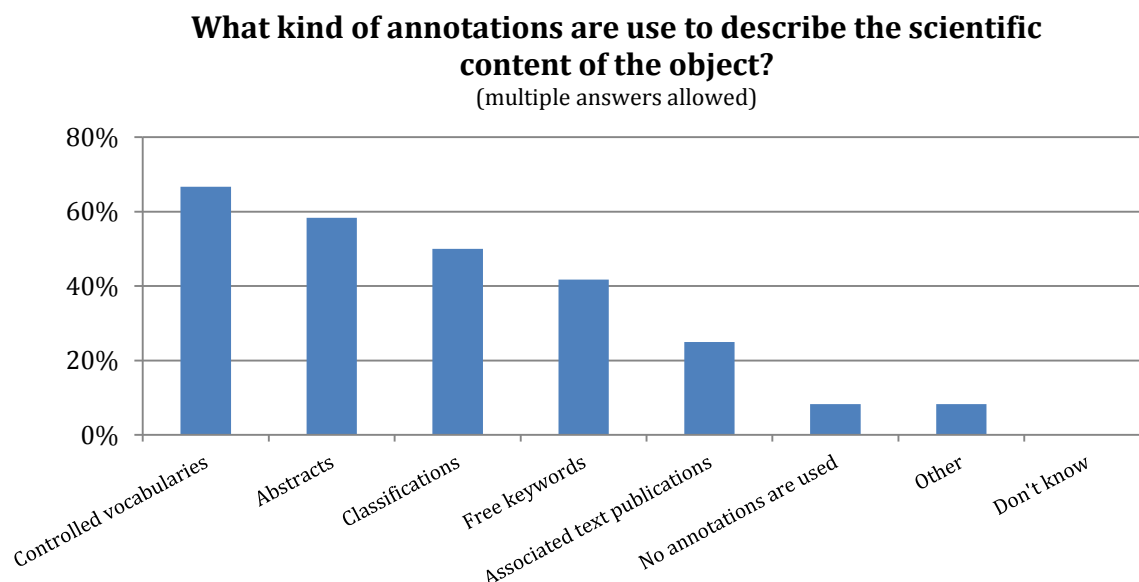
Registration is needed for 67% of the repositories. For 20% the registration is only needed for sub-collections. In some cases modifications of the data (e.g. anonymisation) are necessary to allow access.

There is a general consensus that the accessibility of the data should be restricted as little as possible. Based on this principle a wide and heterogeneous set of legitimate reasons exists for limited access only. Research probing the justification of access restrictions to data in data repositories is a desideratum.

## 3.2 Annotation services

Annotations like standardized metadata or free comments are required to enable re-use of the stored data. 50% of the respondents stated that annotations are "very important" for a possible re-use of data. The chosen metadata standards depend on the data. According to the respondents, in most cases the Dublin Core metadata element set is used to describe the data formally (64%) followed by different ISO standards. Further a variety of standards are used.

Figure 8 shows the kinds of annotation used to describe the content of the object. Controlled vocabularies, abstracts, classifications are the three most frequently named kinds of annotations.



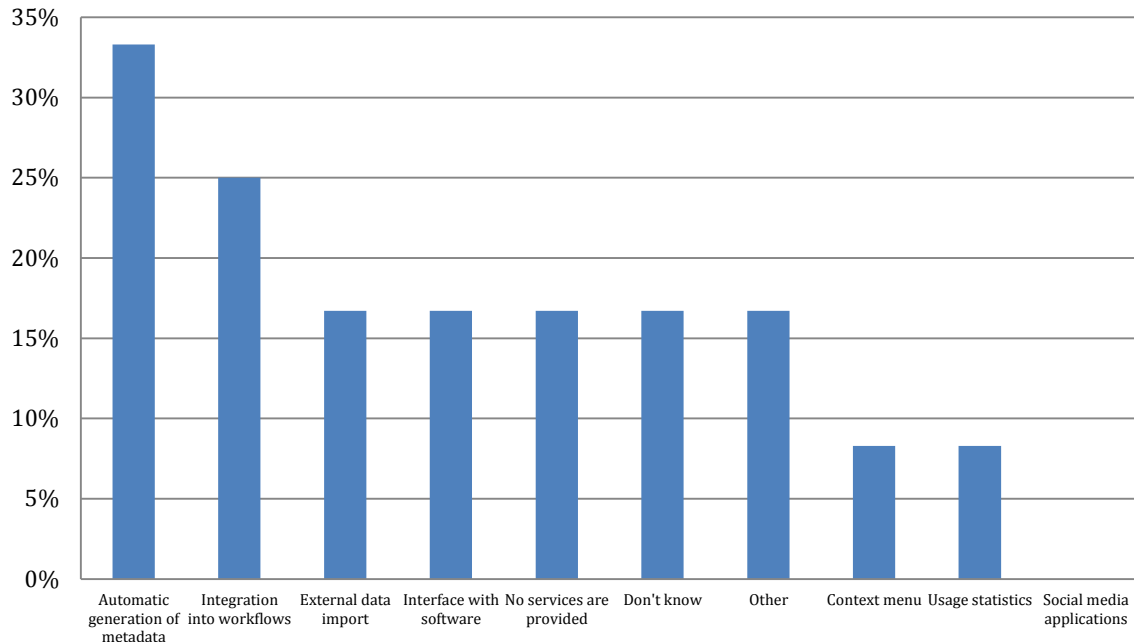
**Figure 8 Used annotations to describe the content of the object, n=12**

In another question, it was asked who usually annotates the data. Multiple answers were permitted. Generally this is done by the data producers (75%) and by the data librarians or data curators (67%).

Interesting are the answers to the question of services for data annotation. Figure 9 shows the various procedures by the members of the APARSEN network. The automatic generation of metadata and the close integration of the data repository into internal workflows of research projects are the two most-mentioned annotation services.

### What services does the repository provide to annotate the data?

(multiple answers allowed)



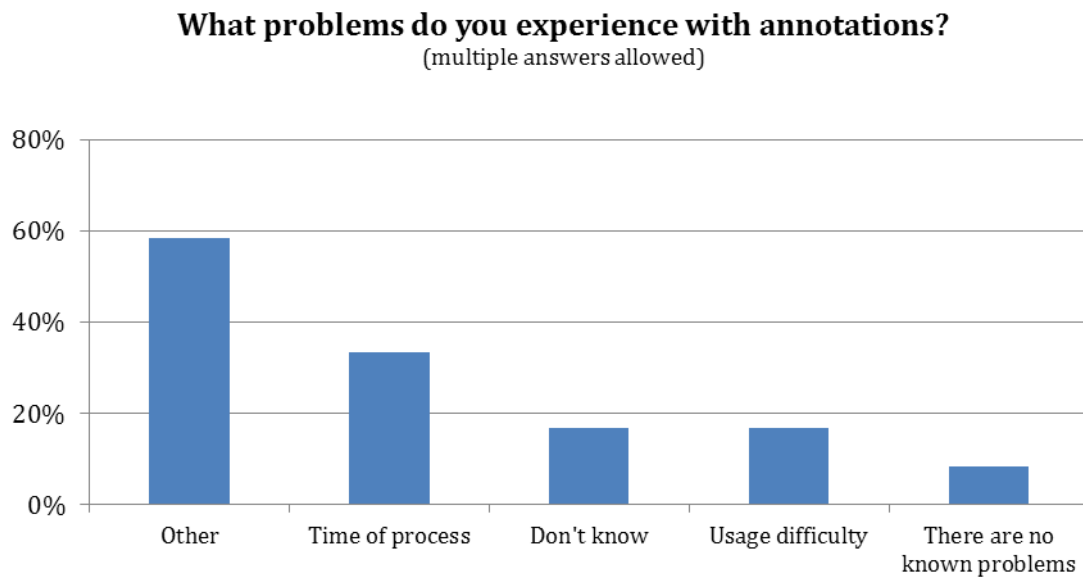
**Figure 9 Annotation services, n=12**

In some cases a detailed description of the data is necessary. The respondents were asked how many fields are used in the typical annotation scheme in use at their data repository. In the majority 5-10 fields are used, in some cases even 50-100.

The survey also queried the challenges of annotation. Figure 8 shows a broad range of specific challenges. In the category "other" the following problem areas were mentioned:

- Poor metadata quality, especially when metadata are imported or generated automatically
- Missing metadata
- Varying names for the same item
- Not all necessary metadata can be delivered by the data provider
- Poor annotations makes data origin and intended (or allowed) use hard to find/understand
- User not identifying the data properly

The variety of answers shows the challenges associated with data annotation.



**Figure 10 Problems with annotations**

Respondents were also asked about the future challenges in the field of annotation. The following upcoming topics were mentioned:

- Availability of annotations as web services
- Combining and harmonizing of different metadata schemes
- Development of an automated service for annotations
- Development of ontologies, to deal with multiple names for the same item
- Enhancing the quality of metadata on the data provider side
- Incentives for annotation

The importance of developing automatic methods to assess the quality of annotation was frequently mentioned.

### 3.3 Reputation

In the area of data repositories there are different dimensions of reputation. For example the reputation of:

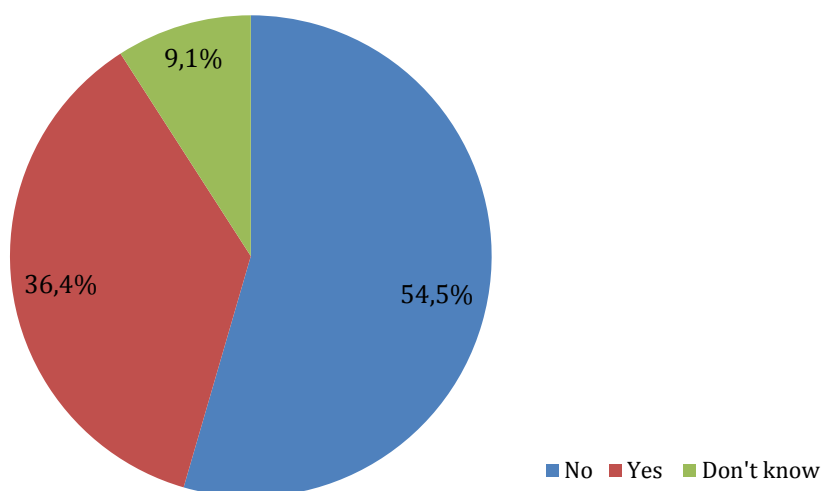
- the repository,
- the host of a repository,
- a stored data set,
- the data producer or
- the reputation of data user.

All these actors (host, repository, data set, data producer and data user) are interdependent. For example: The reputation of a data set affects the reputation of a repository. Further, the reputation of a data producer affects also the reputation of a data set.

The first set of questions in this part of the survey dealt with the reputation of data itself. Figure 11 shows that around 36% of the respondents are tracking the re-use of stored data sets. The following methods are being applied for tracking:

- Analysis of the data repository usage data
- Citation analyses, by using the data sets persistent identifier
- If a registration is required, analysis of the login data

### Do you track the reuse of the data?



**Figure 11 Tracking the re-use of data, n=11**

Suggestions for citation formats are important to facilitate the tracking of data re-use. 54% of the responding APARSEN repositories suggest a citation style to their users, in the case of re-use. 45% do not suggest a citation format. Only 19% of the repositories must be cited in the case of data reuse.

Citation analysis can possibly also help to appraise the relevance of a data repository or a research field and thus help to evaluate justification of its funding.

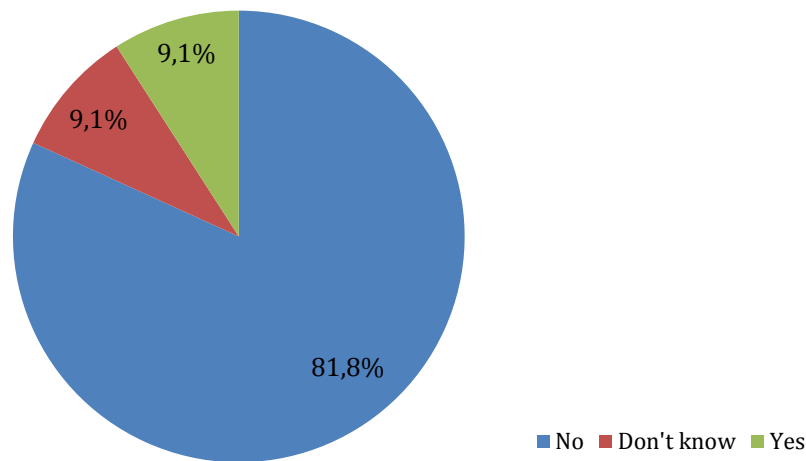
In a further question it was asked if the delivery of data sets is being recorded in an evaluation system. This is mostly not the case (see Figure 12).

It was also asked how methods or tools of reputation can be advanced. The following answers were given:

- Accreditation and certification of repositories
- Clear provenance of data collections
- Development of peer review processes for data sets
- Impact factor for research data repositories (e.g. based on number/relevance of journal publications that rely on data from the respective data repository)
- Improved support for producers and users throughout the data lifecycle.
- Open policies and procedures in the handling of data



### Is the delivery of data sets being recorded in an evaluation system?



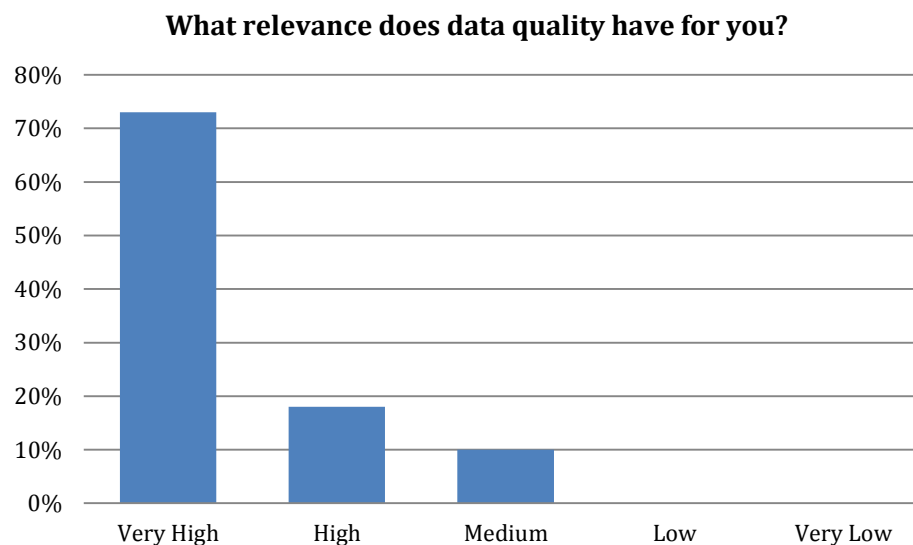
**Figure 12 Reporting of data in an evaluation system, n=11**

### 3.4 Data quality

High quality of data is an essential topic. The quality of a data set is affected by many factors. The individual relevance of these factors depends strongly on research discipline and data type. The respondents were asked how they judge what data can be trusted. The following answers describe the views of the APARSEN members on this topic:

- Availability of clear procedural evidence of best practice throughout the data life cycle
- Availability of detailed provenance information
- Availability of procedures to check the data-integrity.
- Certified data repository
- Checksum
- Data managed by a trusted organization
- Documentation of quality assurance methods
- Provenance and reputation of the data repository
- Quality of metadata
- Reputation of the data producer

Figure 13 illustrates the importance of data quality for the respondents' APARSEN members:



**Figure 13 Relevance of data quality, n=11**

The respondents were also asked by what measures their data repository supports the quality assurance. The following measures were mentioned:

- Business process documentation
- Completeness / Consistency checks
- Data curators technical review (methods, parameters, unit checks, consistency)
- Data management and sharing training
- File format validation
- Metadata checks
- Risk management
- Storage integrity verification
- Tools for annotating quality information

There is a lot of activity in the field of data repository audit and certification. Figure 14 shows the distribution of the various certificates or criteria catalogues in use in the APARSEN network. The Data Seal of Approval (DSA)<sup>110</sup> is the most widespread certificate in the network. DANS, one of the APARSEN partners, has been working on the data seal of approval since 2005. In consultation with various archives and scientific institutions, the first version was distributed within limited circles at the end of 2007. In 2008 the DSA was presented internationally. The DSA can be granted to any data repository that passes the assessment procedure. The results show the variety of certificates or criteria catalogues already in use. In addition to the "well known" some specific certificates or criteria catalogues were mentioned (for example the ICSU World Data System (WDS) membership criteria<sup>111</sup>).

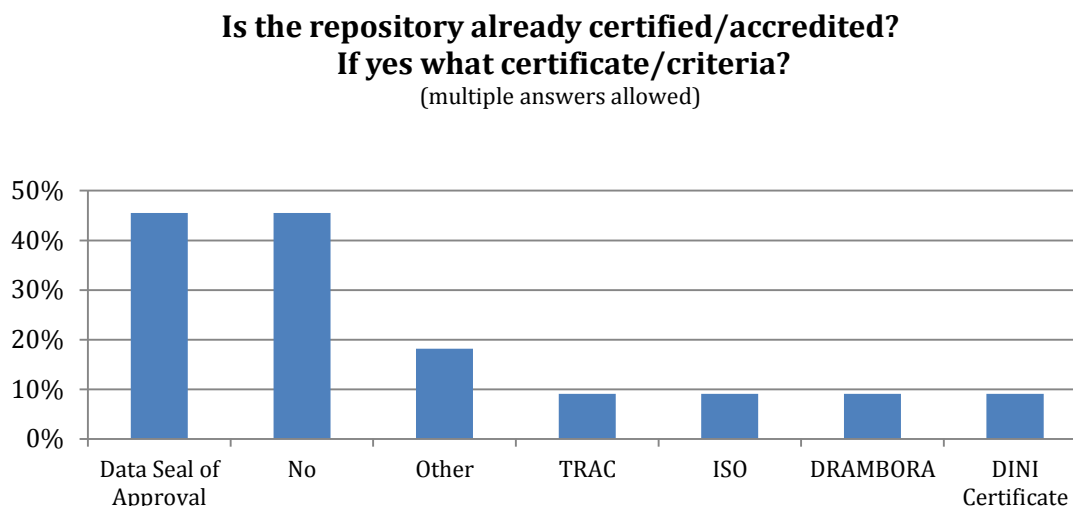
<sup>110</sup> <http://www.datasealofapproval.org/>

<sup>111</sup> <http://www.icsu-wds.org/wds-members/join-icsu-wds/criteria-membership-certification>

At the time of the survey, some partners were still working on (higher levels of) certification of their repositories, e.g. on audit and certification according to

- ISO 16363-DIS 16363112 (therefore only one repository could offer to be already certified)
- DIN 31644113 (not to be confused with DINI)

The activity has been amply documented in APARSEN WP 33<sup>114</sup>, including the relationship between these and their predecessors, such as TRAC.<sup>115</sup>



**Figure 14 Certificates or criteria catalogues in the APARSEN network, n=11**

In a further question respondents were asked about the training measures in data management to enhance data quality. Most of the responding APARSEN members offer trainings for data producers. Some also hold workshops to train their data curators.

<sup>112</sup> Requirements for Bodies Providing Audit and Certification of Candidate Trustworthy Repositories [http://www.iso.org/iso/iso\\_catalogue/catalogue\\_tc/catalogue\\_detail.htm?csnumber=56510](http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=56510)

<sup>113</sup> <http://www.nabd.din.de/projekte/DIN+31644/de/117956308.html>

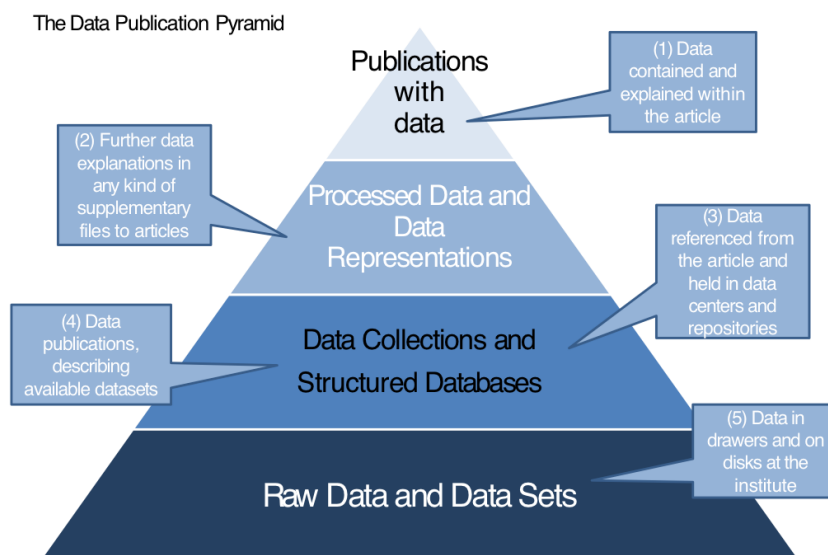
<sup>114</sup> APARSEN D33.1 B: Report on Peer Review of Digital Repositories <http://www.alliancepermanentaccess.org/wp-content/plugins/download-monitor/download.php?id=Report+on+Peer+Review+of+Digital+Repositories>

<sup>115</sup> <http://www.nabd.din.de/projekte/DIN+31644/de/117956308.html>  
<http://wiki.digitalrepositoryauditandcertification.org>  
<http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying-0>  
<http://www.icsu-wds.org/wds-members/join-icsu-wds/criteria-membership-certification>  
[http://www.iso.org/iso/iso\\_catalogue/catalogue\\_tc/catalogue\\_detail.htm?csnumber=56510](http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=56510)  
[http://www.iso.org/iso/catalogue\\_detail.htm?csnumber=57950](http://www.iso.org/iso/catalogue_detail.htm?csnumber=57950)

## 4 Analysis and upcoming research strategies

### 4.1 Critical analysis

Growth in the digitization of science is opening up a wide range of opportunities for scientists. Illustration 14 shows the relation between publications and research data. The exchange of scientific results independent of time and location, collaboration in virtual research environments or the inclusion of laymen in the scientific process of cognition within the scope of so-called “citizen science” are just some examples of the potential of digital science. New perspectives have also emerged for reputation assurance of scientific information. Comment and assessment functions as well as new processes for checking plagiarism are examples of the new opportunities which are being incorporated in daily scientific work increasingly.



**Illustration 14 Data Publication Pyramid<sup>116</sup>**

In addition to the various opportunities provided, there is also a wide range of challenges. As a result of digitization, scientific disciplines are faced with the task of organizing and permanently maintaining a fast growing volume of digital research data. To enable excellent science it is essential to ensure lasting access to these digital information items and to be assured about its quality and usefulness.

Therefore, quality assurance of scientific information is an essential precondition and preserving quality-related information has to be an integral component of digital long-term archiving.

<sup>116</sup> Reilly S., Schallier W., Schrimpf S., Smit E., Wilkinson, M., “Report on integration of data and publications” (2011) [www.alliancepermanentaccess.org/wp-content/plugins/download-monitor/download.php?id=ODE+Report+on+Integration+of+Data+and+Publications](http://www.alliancepermanentaccess.org/wp-content/plugins/download-monitor/download.php?id=ODE+Report+on+Integration+of+Data+and+Publications)

So far a semi-classical view of the infrastructures of science and their properties, as depicted in illustration 9, which was developed from viewpoints collected in the publishers' and librarians' communities by the ODE project, holds.

This view is also reinforced by the APARSEN "Report on peer review of research data in scholarly communication"<sup>117</sup> which documents and categorises ideas, attitudes, developments and discussion concerning quality assurance of research data. The focus is on action taken by scientists, e-infrastructure providers and scientific journals. However, potential cracks begin to appear (p.17).

In this report however, desk research (ch. 1), a snapshot of research topics at APARSEN partners (ch. 2) and a survey of their practises (ch. 3) in handling annotation, reputation and quality show a staggering array of challenges to be addressed.

In particular it addresses those parts of the research process *not* covered by (semi-)classical publishing, such as treating the development of data from raw to primary resource, which in the case of remote sensing data can be perceived as a stand-alone resource – data as scientific results on their own.

At the other end of the interconnected-ness spectrum stands the example of preserving complex workflow environments – which certainly cannot all be wrapped into "executable papers". It raises the question, whether the whole research environment can or needs to be preserve-able and be preserved.

Surrounding all three scenarios – semi-classical, data only and full inclusion of computing - examples and ideas bubble up to greater or lesser degree which show how annotations – including and beyond metadata - are needed for various reasons: Quality assurance, conveying reputation or just making data more useful and useable. It should not really surprise anyone that this theme is so much more important in the context of data: Most of the (human readable) hints, additions or connections they provide can and should already be carried by the main text of a journal article itself, without resorting to external annotations. This is simply not possible in the case of binary data objects, say, a recording of water content along a sediment core.

Although some of the impacts on preservation described, such as preservation and migration of explicit "recommendations" are applicable to journal articles as well, their significance is much higher in the case of data: In publishing journals online, not too much has changed, in particular regarding reputation and quality assurance – nor need it be changed, rapidly, as long as there is no better alternative. In contrast to this, better practises and infrastructural support for data management are needed urgently.

The heterogeneity, unconnected-ness and perhaps even contradictory nature of the evidence and approaches displayed in ch. 1 to 3 need to be seen as evidence for the vastness of the field, its immaturity – in other words: The need of more research, as well at the conceptual level, perhaps cutting some Gordian knots, as at the detailed, technical level, shows clearly.

## 4.2 Research strategies

Based on internal consultation the following research areas on annotation, reputation and data quality were named as priorities by APARSEN members.

In the area of social sciences and humanities the approach to use workflow management systems is perhaps less strong than in science disciplines, but even here there is certainly a trend towards combining data with applications for accessing and analysing them. As initially such applications are

---

<sup>117</sup> Pampel, H., Pfeiffenberger, H., Schäfer, A., Smit, E., Pröll, S., & Bruch, C. (2012). Report on Peer Review of Research Data in Scholarly Communication. Retrieved from <http://epic.awi.de/30353/>

typically custom-built, it makes sense to preserve the data and the application in conjunction<sup>118</sup>. One of the questions to answer is what “preserving the application” entails:

- Does it imply to keep a live environment up to date (for how long?) and/or should the source code or the application’s significant properties be archived<sup>119</sup>?

Several partners are involved in “experiments” to develop peer review and peer recommendation of data. These reviews are made a posteriori, sometimes many years after the dataset has been deposited. The review processes connect the aspects of reputation – by providing feedback to the original researcher(s) –, data quality – by rating several aspects of the data and the metadata – and annotation, since the reviews enhance the annotation (metadata) of the dataset and support interested researchers in deciding whether a dataset is of relevance for their research.

- These experiments deserve a range of follow up research, perhaps regarding their effectiveness and efficiency, but definitely regarding their scalability and integration into the research process.

Finally, at the highest conceptual level, it is an open question

- whether general or broader disciplinary rules or catalogues can be formulated as to which (static) auxiliary information – annotation – needs to be and can be preserved for how long (and how access has to be granted) in order to keep the scientific process healthy.

Most of these topics would have to be developed in close cooperation between data technologists, sociologists of science and experienced scientists from many domains – the later perhaps best to be found in learned societies.

Regarding research at the “details” level, the authors are convinced that (multiple) clues can be found on almost every page of chapters 1 to 3. However, some aggregation – at current conceptual development – is being tried in the following subsections.

#### 4.2.1 Annotation and annotation services

The accurate and detailed description of data is indispensable for the future re-use. The further development of annotation services is necessary to enable the chances of digital science. Research topics to be addressed in future include:

- Innovative scenarios for the re-use of annotations
- Development of user friendly tools for annotation
- Development of social annotation services
- Semantic web services for annotation
- Interoperability of metadata
- Aggregation services for different metadata types
- Migration services for metadata
- Developments of authoritative (and dominant) registries of names, concepts, etc.

Further:

- Promoting open metadata initiatives.

---

<sup>118</sup> A striking precedent has been set recently by ENCODE (the so called Human Genome Project 2.0): It provided their analysis environment as a virtual machine for execution in a cloud. [www.isgtw.org/feature/human-genome-project-20](http://www.isgtw.org/feature/human-genome-project-20)

<sup>119</sup> Matthews, B., B. McIlwrath, D. Giaretta, E. Conway: The significant properties of software: A study. Science and Technology Facilities Council, March 2008. Retrieval via <http://bit.ly/eF7yNv>



- Repeating surveys of practice every few (5?) years to track changes (semi-)quantitatively

## 4.2.2 Reputation

Reputation is an important incentive that helps fill repositories and thus to provide access and re-use of scientific data, in the first place. The APARSEN members identified the following topics to promote the reputation of data and data producers.

- Promotion of data citation standards
- Developing standards for assigning and maintaining identifiers to digital objects
- Development of bibliometric methods for data citation
- Development of other statistical methods for data usage<sup>120</sup>
- Derivation of new metrics and incentives from reputation
- Methods of evaluation of reputation information in automated workflows and algorithms
- Rules and practises to establish reputation of repositories (e.g. certification)
- Dynamics of transfer of reputation from creators to data and vice versa; from repositories to data and vice versa

## 4.2.3 Data quality

Quality of data is of high importance for the APARSEN members. Various measures can be taken to improve the quality of data. The following research needs are detected:

- Standardization of review methods and processes for data and encoding and classification of their results
- Improvement and standardization of quality assurance methods for data production and data infrastructures
- Development of best practise curricula and training methods for data producers and data curators

---

<sup>120</sup> See e.g. the guidelines from Knowledge Exchange and the SURF SURE project  
<http://wiki.surf.nl/display/standards/KE+Usage+Statistics+Guidelines>

## 5 Conclusions and Recommendations

The topics of this report – in particular reputation and quality – belong to the domain of Trust. Each element and the whole fabric of a future data management and preservation infrastructure must support the production and identification of trustworthy data and enable the discarding or separating of untrustworthy. Failing that, it would not only fail to attract continued use, but also bury valuable research resources, and thus results of uncounted researchers, under irrelevant or misleading junk.

The report is thus considered to be potentially relevant to (almost) all other work packages of APARSEN, from identifiers to business models.

The field of auxiliary information to research data is – technically and conceptually – much wider, much younger and less developed than the comparable field for textual information. Each report such as this can only scratch at the surface or at best identify *some* of the key questions, which will arise.

Thus, putting up individual findings and some suggested research topics for discussion (or, hopefully, for immediate uptake), the authors suggest this collection be studied and extended and its conclusions further broadened and honed in the organizational context of the APARSEN consortium and the upcoming virtual centre of excellence and its facilities for discussion and dissemination, becoming a living document.

Observing that current trends in research funding for data infrastructures care much about scale and volume, it is to be hoped that concepts to establish trust can still be introduced in time. A large number of APARSEN members are well positioned to do this, with an effect.

## 6 List of figures

Figure 1 Disciplines of stored data, n = 15 .....	43
Figure 2 Data types, n = 16 .....	44
Figure 3 Funding of the repositories - now and in the future, n = 15 .....	45
Figure 4 Amount of stored data, n = 16 .....	45
Figure 5 Web hits per month, n = 16 .....	46
Figure 6 Actors in ingest process, n = 16 .....	47
Figure 7 Storage - voluntary or mandatory, n=16 .....	48
Figure 8 Used annotations to describe the content of the object, n=12 .....	49
Figure 9 Annotation services, n=12 .....	50
Figure 10 Problems with annotations .....	51
Figure 11 Tracking the re-use of data, n=11 .....	52
Figure 12 Reporting of data in an evaluation system, n=11 .....	53
Figure 13 Relevance of data quality, n=11 .....	54
Figure 14 Certificates or criteria catalogues in the APARSEN network, n=11 .....	55

## 7 List of tables

Table 1 The four levels of research data in HEP in order of increasing complexity .....	26
---	----

## 8 List of Illustrations

Illustration 1 Collaborative Data Infrastructure .....	8
Illustration 2 Scientific workflow modelled in the Taverna Workflow engine .....	22
Illustration 3 Header of a data table on the data repository site .....	27
Illustration 4 Header of a data table on INSPIRE .....	27
Illustration 5 Examples for schema/ontology evolution (a) .....	29
Illustration 6 Examples for schema/ontology evolution (b) .....	29
Illustration 7 RIMQA - Start screen .....	30
Illustration 8 RIMQA - Recommendation of possible refinements .....	31
Illustration 9 RIMQA - Recommendation of possible refinements for one particular object .....	31
Illustration 10 Example of a data citation .....	33
Illustration 11 Assessment of the dataset "De steentijd van Nederland" .....	34

Illustration 12 Quality Flag Scheme of the COSYNA project .....	36
Illustration 13 Data Level - Overview of the COSYNA project .....	37
Illustration 14 Data Publication Pyramid .....	56

## 9 References

- Albani, M., Beruti, V., Duplaa, M., Giguere, C., Velarde, C., Mikusch, E., Serra, M., et al. (2010). Long Term Data Preservation of Earth Observation Space Data. European LTDP Common Guidelines. Issue 1.1. (M. Albani, Ed.). Retrieved from [http://earth.esa.int/gscb/ltdp/EuropeanLTDPCommonGuidelines\\_Issue1.1.pdf](http://earth.esa.int/gscb/ltdp/EuropeanLTDPCommonGuidelines_Issue1.1.pdf)
- Bruch, A. A., Uhl, D., & Mosbrugger, V. (2007). Miocene climate in Europe — Patterns and evolution A first synthesis of NECLIME. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 253(1-2), 1-7. doi:10.1016/j.palaeo.2007.03.030
- Breure, L., Voorbij, H., Hoogerwerf, M. (2011) Rich internet publications: Show what you tell. *Journal Of Digital Information*, Vol 12(Nr. 1)
- Credit where credit is overdue. (2009). *Nature biotechnology*, 27(7), 579. doi:10.1038/nbt0709-579
- Curcin, V., Ghanem, M.. Scientific workflow systems - can one size fit it all?. (2008). *Biomedical Engineering Conference, CIBEC 2008*.
- Data Archiving and Networked Services. (2011). Data Reviews. Peer-reviewed research data. Retrieved from <http://www.dans.knaw.nl/en/content/categorieen/publicaties/dans-studies-digital-archiving-5>
- David, K., Santos, E., Mates, P., Vo, H.T., Bonnet, P., Bauer, B., Surer, B., Troyer, M., Williams, D., Tohline, J., Freire, J., Silva, C. (2011). A provenance-based infrastructure to support the life cycle of executable papers. *Proceedings of the International Conference on Computational Science, ICCS 2011*
- De Roure, D., Belhajjame, K., Missier, P., Gomez-Perez, M., Palma, R., Ruiz, J. (2011) Towards the preservation of scientific workflows. *8th International Conference on Preservation of Digital Objects iPRES 2011*.
- DPHEP Study Group. (2009). Data Preservation in High-Energy Physics. Retrieved from <http://arxiv.org/abs/0912.0255>
- DPHEP Study Group. (2012). Status Report of the DPHEP Study Group: Towards a Global Effort for Sustainable Data Preservation in High Energy Physics. Retrieved from <http://arxiv.org/abs/1205.4667>
- Klump, J., Bertelmann, R., Brase, J., Diepenbroek, M., Grobe, H., Höck, H., Lautenschlager, M., et al. (2006). Data publication in the open access initiative. *Data Science Journal*, 5, 79-83. doi:10.2481/dsj.5.79
- Kuipers, T., & Van der Hoeven, J. (2009). Insight into digital preservation of research output in Europe. Survey Report. Framework. Retrieved from [http://www.parse-insight.eu/downloads/PARSE-Insight\\_D3-4\\_SurveyReport\\_final\\_hq.pdf](http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf)
- Lin, C., Lu, S., Fei, X., Chebotko, A., Pai, D., Lai, Z., Fotouhi, F., Hua, J., (2009). A Reference Architecture for Scientific Workflow Management Systems and the VIEW SOA Solution. *IEEE Transactions on Services Computing*, 1(2), 79-92. doi 10.1109/TSC.2009.4

Mayer, R., Rauber, A., Neumann, M.A., Thomson, J., Antunes, G. (2012). Preserving scientific processes from design to publication. Proceedings of the 15th International Conference on Theory and Practice of Digital Libraries (TPDL 2012)

Pampel, H., Pfeiffenberger, H., Schäfer, A., Smit, E., Pröll, S. & Bruch, C. (2012). Report on Peer Review of Research Data in Scholarly Communication. Retrieved from <http://epic.awi.de/30353/>

Pfeiffenberger, H., & Carlson, D. (2011). "Earth System Science Data" (ESSD) - A Peer Reviewed Journal for Publication of Data. D-Lib Magazine, 17(1/2). doi:10.1045/january2011-pfeiffenberger

QA4EO Task Team. (2010). A Quality Assurance Framework for Earth Observation. Principles. Version 4.0. Retrieved from [http://qa4eo.org/docs/QA4EO\\_Principles\\_v4.0.pdf](http://qa4eo.org/docs/QA4EO_Principles_v4.0.pdf)

Reilly, S., Schallier, W., Schrimpf, S., Smit, E., & Wilkinson, M. (2011). Report on Integration of Data and Publications. Retrieved from <http://ode-project.eu/outputs>

Sierman, B., Schmidt, B., Ludwig, J. (2009) Enhanced Publications : Linking Publications and Research Data in Digital Repositories. Surf EU-Driver. Amsterdam University Press, Amsterdam.

South, D. (2011). Data Preservation in High Energy Physics. Retrieved from <http://arxiv.org/abs/1101.3186>

Southan, C., & Cameron, G. (2009). Database Provider Survey. Report for ELXIR Work Package 2 (30th ed.). Retrieved from [http://www.elixir-europe.org/bcms/elixir/Documents/reports/WP2\\_Annex-Provider\\_Survey\\_Report.pdf](http://www.elixir-europe.org/bcms/elixir/Documents/reports/WP2_Annex-Provider_Survey_Report.pdf)

## 10 Annex: List of questions

### Internal Survey: "Annotation, Reputation and Data Quality"

#### 1. Please list your affiliation:

- Airbus Operations SAS (Airbus Operations)
- Alliance Permanent Access (APA)
- Austrian National Library (ONB)
- CINES
- CINI (Consorzio Interuniversitario Nazionale per l'Informatica)
- CSC - Tieteen tietotekniikan keskus Oy (CSC)
- Deutsche Nationalbibliothek (DNB) – German National Library
- Digital Preservation Coalition (DPC)
- European Organisation for Nuclear Research (CERN)
- European Space Agency (ESA)
- Fondazione Rinascimento Digitale (FRD)
- Forschungsinstitut für Telekommunikation (FTK)
- FORTH Foundation for Research and Technology-Hellas (FORTH)
- Globale Informationstechnik GmbH (GLOBIT)
- Helmholtz Association
- IBM Israel, Science and Technology Ltd
- InConTec GmbH (ICT)
- INMARK Estudios y Estrategias, S.A (INMARK)
- International Association of Scientific, Technical and Medical Publishers (STM)
- KNAW-DANS, Data Archiving and Networked Services
- Koninklijke Bibliotheek (KB)
- Luleå University of Technology (LTU)
- Microsoft Research Limited (MRL)
- Philips Consumer Lifestyle (PCL)
- Science and Technology Facilities Council (STFC)
- Secure Business Austria (SBA)
- Space Research Institute of the Russian Academy of Sciences (IKI RAN)
- Tessella
- The British Library (BL)



- The Stichting LIBER Foundation
- University of Essex, UK Data Archive (UKDA)
- University of Patras, Library & Information Center (LIC), GREECE (UPAT)
- University of Trento (UNITN)
- Other (please specify) :

**2. Please list the name and the address of the data repository:**

- Name:
- Address [URI], if possible:

**3. Please indicate what type of data is stored in the repository (multiple answers are possible):**

- Standard office documents (text documents, spread sheets, presentations)
- Network based data (web sites, email, chat history, etc.)
- Databases (DBASE, MS Access, Oracle, MySQL, etc.)
- Images (JPEG, JPEG2000, GIF, TIF, PNG, SVG, etc.)
- Structured graphics (CAD, CAM, 3D, VRML, etc.)
- Audio-visual (multimedia) data (WAVE, MP3, MP4, Flash, etc.)
- Scientific and statistical data formats (SPSS, FITS, GIS, etc.)
- Raw data (device specific output)
- Plain text (TXT in various encodings)
- Structured text (XML, SGML, etc.)
- Archived data (ZIP, RAR, JAR, etc.)
- Software applications (modelling tools, editors, IDE, compilers, etc.)
- Source code (scripting, Java, C, C++, Fortran, etc.)
- Configuration data (parameter settings, logs, library files)
- Other (please specify):

**4. What discipline can the data be attributed to? (multiple answers are possible)**

- Humanities
- Social Sciences
- Life Sciences
- Natural Sciences
- Engineering Sciences
- Customer Research

- Other (please specify):

**5. Which of these categories can the stored data be attributed to? (multiple answers are possible)**

- Research data
- Governmental data
- Cultural data
- Internal company data
- Other (please specify):

**6. Who submits the data to the repository?**

- Data producer
- Data librarian / data curator
- Mixed
- Other (please specify):

**7. Is the storage in the repository for the target group voluntary or mandatory?**

- Voluntary
- Mandatory
- Don't know
- Mixed (please specify in which sense):

**8. How can the stored data be accessed?**

- Access is restricted
- Access is open
- Mixed
- Don't know

**9. If the access is restricted, please specify:**

- Data is only accessible for a specific research discipline
- Data is only accessible for a specific research group
- Data is only accessible for a fee
- Access to the data is temporarily restricted
- Mixed
- Other (please specify):

**10. Do repository users need to register?**

- Yes
- No
- Mixed
- Don't know

**11. Must the repository be cited in the case of data reuse?**

- Yes
- No
- Mixed
- Don't know

**12. Are changes necessary to open the access to the data (e. g. anonymisation)?**

- Yes
- No
- Mixed
- Don't know

**13. What amount of data is stored in the repository? Please estimate also the volume in 2 and 5 years.**

- 0 MB
- 1-100 MB
- 100 MB- 1 GB
- 1 GB-1 TB
- 1 TB-1 PB
- 1 PB-1 EB
- 1EB
- Don't know

**14. Where you can, please supply web hits per-month (excluding web-crawling):**

- 0 - 500
- 500 - 1000
- 1 K - 5 K
- 5 K - 10 K

- 10 K - 50 K
- 50 K - 100 K
- 100 K - 500 K
- 500 K - 1 M
- 1 - 5 M
- 5 - 10 M
- > 10 M
- Don't know

**15. Where you can, please supply the number of unique users per-month:**

- 0 - 500
- 500 - 1000
- 1 K - 5 K
- 5 K - 10 K
- 10 K - 50 K
- 50 K - 100 K
- 100 K - 500 K
- 500 K - 1 M
- 1 - 5 M
- 5 - 10 M
- >10 M
- Don't know

**16. What type of funding does your repository have? (If mixed, please tick multiple boxes)**

- Funding outside Europe
- Institutional
- National grant
- Rolling funding
- European funding
- Intermittent
- Commercial
- No formally specified funding (e.g. the repository was a by-product of a research project)
- Don't know
- Other (please specify):

## **17. About the funding:**

Is it a problem for you now?

- Yes
- No
- Don't know

Will it be a problem in 5 years time?

- Yes
- No
- Don't know

Will it be a problem in 10+ years time?

- Yes
- No
- Don't know

## **18. Does the repository have any of the following preservation strategies in place? (multiple answers are possible)**

- Migration (periodic conversions of file formats to popular formats of today)
- Normalisation (conversion of all publications to one standardized file format sustainable over time)
- Emulation (no conversions of the original publication but capturing the original context)
- Outsourced to a third party service
- No preservation strategies in place
- Don't know
- Other (please specify):

## **19. What kind of measures do you undertake to guarantee for the sustainability of your repository?**

- No measures are taken
- Don't know
- The following measures are taken:

## **20. Do you arrange training for/give advice for repository users to ensure good practice when they submit data on the repository?**

- Yes
- No

- Don't know

**21. If your institution has developed and hosts multiple repositories please give the total:**

- 1-5
- 5-10
- 10-15
- 15-20
- 25-30
- 35-40
- 45-50
- More
- Don't know

**22. What kind of standards are used to describe the data formally? (multiple answers are possible)**

- Data Documentation Initiative (DDI)
- Dublin Core (DC)
- ISO
- MAB (Automated Library Exchange Format)
- Machine-Readable Cataloging (MARC)
- Metadata Object Description Schema (MODS)
- Text Encoding Initiative (TEI)
- No standards are used
- Don't know
- Other (please specify):

**23. If ISO, please specify:**

- (text field)

**24. What kind of annotations are use to describe the scientific content of the object? (multiple answers are possible)**

- Free keywords
- Controlled vocabularies
- Classifications (e.g. DDC)
- Abstracts

- Associated text publications (e.g. scholarly journals)
- No annotations are used
- Don't know
- Other (please specify):

**25. What services does the repository provide to annotate the data? (multiple answers are possible)**

- Automatic generation of metadata
- Context menu
- External data import
- Integration into workflows
- Interface with software
- Usage statistics
- Social media applications (Facebook, Twitter, etc.)
- No services are provided
- Don't know
- Other (please specify):

**26. Who annotated the data? (multiple answers are possible)**

- Data producer
- Data librarian / data curator
- Data user
- Mixed
- No one
- Don't know
- Other (please specify):

**27. Are the data producers named in the annotation?**

- Yes
- No
- Mixed
- Don't know



**28. Can the metadata be commented by the users?**

- Yes
- No
- Don't know

**29. How many fields per annotation have to be filled in for a typical data set?**

- 1-5
- 5-10
- 10-15
- 15-20
- 20-25
- 30-35
- 35-40
- 45-50
- 50-100
- More than 100
- No annotations are used
- Don't know

**30. Percentage of annotation fields usually filled for your typical data set?**

- No annotations are used
- Don't know
- Please estimate percentage:

**31. Are you concerned with the quality of the annotations?**

- No
- Yes
- Don't know
- Explanations are welcome:

**32. What problems do you experience with annotations? (multiple answers are possible)**

- Usage difficulty
- Time of process
- There are no known problems

- Don't know
- Other (please describe):

**33. Evaluate the importance of annotation for a possible re-use of the stored data:**

- Extremely important
- Very important
- Moderately important
- Slightly important
- Not at all important

**34. Are you concerned with the quality of research data when migrated to newer systems?**

- Yes
- No
- Don't know
- Explanations are welcome:

**35. Where do you see the future challenges in the field of annotation?**

- (text field)

**36. Do you track the reuse of the data?**

- No
- Don't know
- Yes (please specify):

**37. Is there a suggested citation style for data?**

- Yes
- No
- Don't know

**38. Are there incentives to store data in the repository?**

- No
- Don't know
- Yes (please specify):

**39. Is the delivery of data sets being recorded in an evaluation system?**

- No
- Don't know
- Yes (please specify):

**40. Is the repository already certified/accredited? If yes what certificate/criteria? (multiple answers are possible)**

- Audit and Certification of Trustworthy Digital Repositories (RAC)
- Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC)
- ISO
- Data Seal of Approval
- Digital Repository Audit Method Based on Risk Assessment (DRAMBORA)
- DIN 31644
- DINI-Certificate Document and Publication Services
- Nestor Catalogue of Criteria for Trusted Digital Repositories
- Don't know
- Other (please specify):

**41. If ISO, please specify:**

- (text field)

**42. Is there a certification or accreditation procedure underway for your repository, if so which ones? (multiple answers are possible)**

- Audit and Certification of Trustworthy Digital Repositories (RAC)
- Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC)
- ISO
- Data Seal of Approval
- Digital Repository Audit Method Based on Risk Assessment (DRAMBORA)
- DIN 31644
- DINI-Certificate Document and Publication Services
- Nestor Catalogue of Criteria for Trusted Digital Repositories
- Don't know
- Other (please specify):

**43. If ISO, please specify:**

- (text field)

**44. Do you see other methods/tools to build reputation of data producers / data repositories?**

- No
- Don't know
- Yes (please specify):

**45. What relevance does data quality have for you?**

- Very High
- High
- Medium
- Low
- Very Low

**46. What measures are being undertaken for quality assurance?**

- No measures are taken
- Don't know
- Please specify the measures:

**47. By which measures the repository supports the quality assurance?**

- No measures are taken
- Don't know
- Please specify the measures:

**48. What measures are being made to ensure that data are interpretable?**

- No measures are taken
- Don't know
- Please specify the measures:

**49. How do you check that the data are interpretable?**

- This is not checked
- Don't know
- Please specify the measures:

**50. How do you evaluate your current documentation for the re-use of research data?**

- Very Good
- Good
- Moderate
- Bad
- Very Bad

**51. Do you employ or foresee a third party (peer-) review process for your data, if so which one?**

- No
- Don't know
- Yes (please specify):

**52. What kind of training measures in data management do you foresee to enhance data quality?**

- No measures are taken
- Don't know
- Please specify the measures:

**53. What are the future challenges in the context of data quality for you?**

- Don't know
- Please describe the future challenges:

**54. How do you as an expert, judge what data can be trusted?**

- (text field)