APA
RSEN

Alliance Permanent Access to the
Records of Science in Europe Network

trust

is fundamental to the working of society

In particular when it comes to unfamiliar digitally encoded information, especially when it has passed through several hands over a long period of time.

APA
ALLIANCE FOR PERMANENT ACCESS

SEVENTH FRAMEWORK PROGRAMME

"Has it been preserved properly?"

A European Framework for Audit and Certification of Digital Repositories

Trust is fundamental to the working of society – in particular when it comes to unfamiliar digitally encoded information, especially when it has passed through several hands over a long period of time.

- **Has it been preserved properly?**
- **Is it of high quality?**
- **Has it been changed in some way?**
- **Does the pointer get me to the right object?**

APARSEN has collected, evaluated and developed the key answers to these questions.

There has long been a demand for some way to evaluate the ability of repositories to preserve the digitally encoded information with which they are entrusted. Several are becoming available – but how do they fit together? APARSEN answers this by helping to set up the European Framework for Audit and Certification of Digital Repositories[1].

This provides three levels of evaluation for repositories ranging from the Data Seal of Approval (DSA)[2] as an entry point, which requires a few days' effort from the repositories, to the much more detailed formal standards DIN 31644[3] and ISO 16363[4], which require several person months to collect the evidence and take part in the audits. The focus of the DSA is on smaller organisations within the research data domain. The two more formal standards are more demanding but, in some cases, especially where there are higher stakes involved, will provide a greater guarantee of 'trustworthiness'. A system for formal accreditation of auditors is being put in place.

APARSEN has collected details of a number of test audits which were carried out, the problems encountered and the lessons learned. The test audits against ISO 16363 were performed by members of the Primary Test Audit Board (PTAB), a group of individuals with wide-ranging experience of digital repositories. The European repositories were the Data Archiving and Networked Services (DANS) from the Netherlands, UK Data Archive (UKDA), Centre Informatique National de l'Enseignement Supérieur: Département Archivage et Diffusion (CINES-DAD) and in addition, in the USA, the Socioeconomic Data and Applications Center (SEDAC) at the Center for Earth Science Information, the National Space Science Data Center (NSSDC) and the Kentucky Department for Libraries and Archives (KDLA). A test audit against DIN 31644 at the Deutsche Nationalbibliothek (DNB) was performed by members of the nestor Working Group Certification.

Data Centre – Courtesy of Bob West (CC BY-NC-SA 2.0)

Is being audited worthwhile?
This is what the repositories say:

For DNB, the main motivation for undergoing audit and certification was to have their own processes and documentation reviewed, scrutinized, and ideally approved by some external professionals. The preparation for the test audit and certification required a thorough analysis and documentation of the achieved status. Thereby, strengths as well as gaps were revealed, which is already a valuable result. Feedback from the auditors will influence the medium term development directions, especially in areas where the auditors suggested improvements. For the DNB, this knowledge gain is even more important than receiving a certificate to showcase.

The advantage of the (test) audit for DANS was that it sheds a clear light on what the strengths and the weaknesses are in the archiving activities of our institute. It gave us confidence that we are well on our way to fulfil the requirements. As the procedure was not yet formal, and we do not yet pass all the requirements, we do not use our "marks" yet to promote the archive, although we certainly do mention to our (potential) users that we are determined to be among the first officially certified digital archives.

UKDA stated that the comments have proven instructive.

CINES-DAD said that certainly helped them to evaluate the progress made since the previous audits and the relevance of the actions taken over the past couple of years, and was a good experience as a contribution to a standardization process, as CINES [required] a predefined scale for the self-assessment of the different metrics of the audit.

SEDAC stated that the ISO 16363 test audit provided an excellent opportunity for SEDAC to continue assessing its data management policies and procedures to identify opportunities for improvement.

Very importantly the formal standard based audits identify priority areas in need of improvement in the repositories' preservation activities, and the impact may be seen in the following statements from them:

*The **DNB** said: we will integrate the test audit results into its short and medium term digital preservation development strategy. Two concrete results were that the DNB will have to document more thoroughly its policy decisions and will have to reinforce its internal Quality Assurance.*

***DANS** said: we have taken the recommendations from the test audit as a primary guideline in the further development of our procedures and technical adaptations of our archive. The test audit gave clear indications where we could improve the trustworthiness of our archive.*

***UKDA** said: implementation of non-contentious recommendations have been undertaken.*

***CINES-DAD** said: the observations and report produced by the test audit team certainly validated the procedures put in place and the willingness for transparency, although some requirements for clarification of the documentation were identified and have been fixed since then; additional action plans have also been worked out to address the few metrics which were not satisfied, and progress will be monitored regularly.*

***SEDAC** said: the recommendations received from the test audit are important inputs into SEDAC's efforts to improve its capabilities and practices for data preservation and stewardship in collaboration with the Columbia University Libraries.*

For more information see the full report from the **APA/APARSEN** web site.

If you want to have your repository audited or just to find out more about the process:
http://trusteddigitalrepository.eu
or begin the ISO 16363 self-audit using the spreadsheet at:
http://www.iso16363.org/preparing-for-an-audit/

Key points…

A European Framework has been created to bring together a consistent set of increasingly challenging audit processes.

The three levels are:

- **"Basic" level**
  monitored self-audit using the Data Seal of Approval
- **"Extended" level**
  monitored self-audits using the extensive the DIN 31644 or ISO 16363 standards
- **"Formal" level**
  a full ISO or DIN audit by external accredited auditors

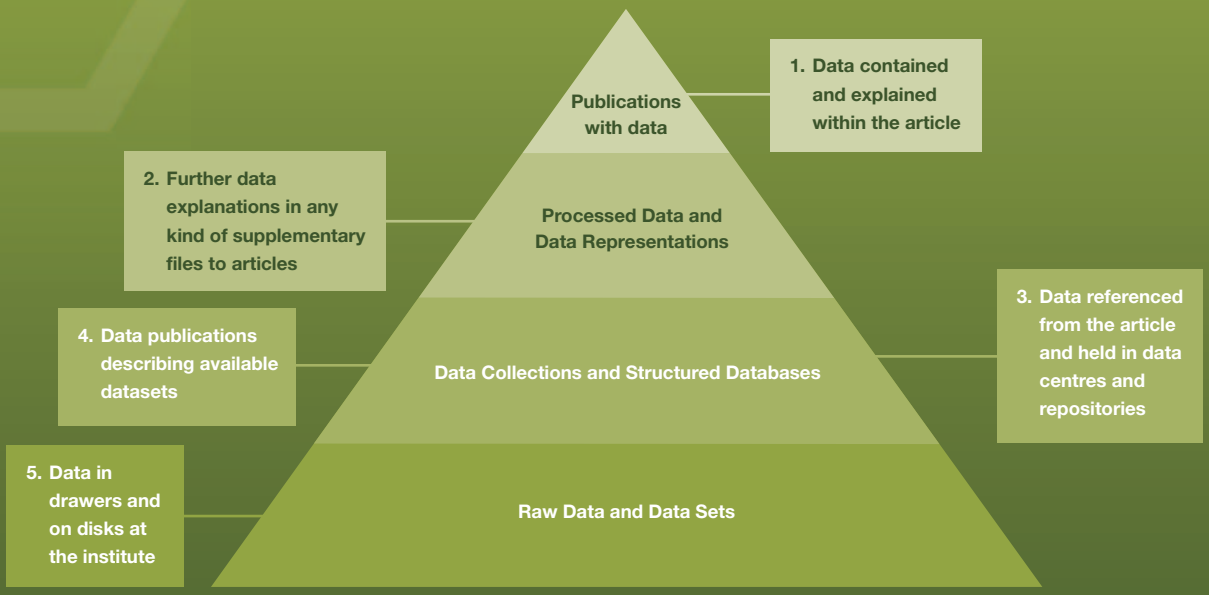
"Is it of good quality?"

Annotation, reputation and data quality

Scientific progress is based on high quality information. The term, quality, is defined in the Academic Press Dictionary of Science and Technology as follows: *"[…] an essential or distinctive characteristic of property of a thing […]"*.[6] The metaphor *"standing on the shoulders of giants",* which vividly describes the scientific cognitive process clearly shows that new findings are always based on statements already published.[7] Access to information of which the quality is assured is therefore a precondition for scientific excellence—the figure below illustrates the dependencies on lower levels.

Growth in digital science is opening up a wide range of opportunities for scientists. Figure 1 illustrates 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 within the scope of so-called "citizen science" are just some examples of the potential of digital science. New perspectives have also emerged for quality 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 increasingly incorporated in daily scientific work.

Figure 1:
The "Data Publications Pyramid"
(from the ODE[8] project)



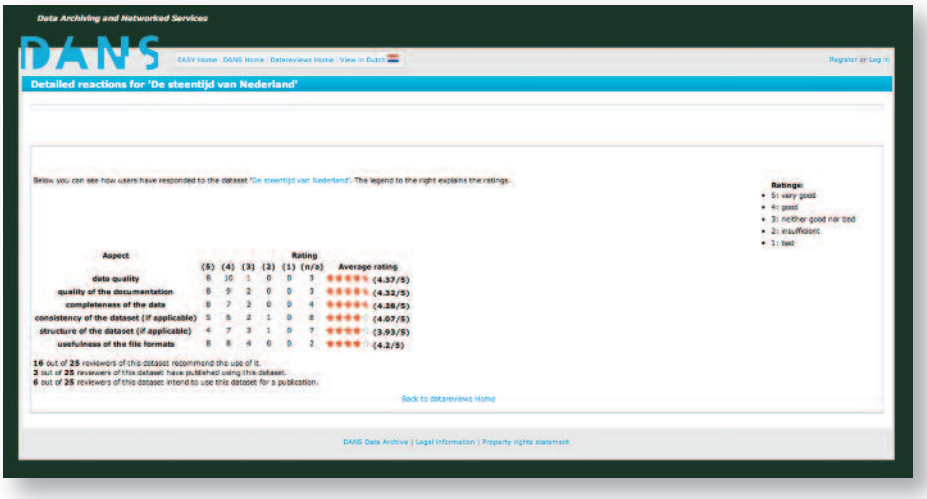
Peer review of data is vital but is much more difficult than for articles. How can it be done?

In addition to the various opportunities provided, there is also a wide range of challenges. As a result of digitization, scientific disciplines in particular 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 this digital data.

Quality assurance of scientific information is an essential precondition and an integral component of digital long-term archiving and is the subject of an APARSEN Report on peer review of research data in scholarly communication 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. Future fields of research are to be described based on this work.

An example of taking advantage of scientists' evaluations is shown below.

Figure 2:
Assessment of the dataset "De steentijd van Nederland" (the stone age in the Netherlands)



Key points…

The studies show a uniform picture of the perspective of scientists to the peer review of scientific data:

- Scientists recognize that accessibility of data is a precondition for peer review of it
- In principle, reviewers and editors find it preferable for data to be peer reviewed but many reservations exist about its feasibility
- Scientists fear that reviewing data in the course of the peer review process is not practical due to the amount of work and time involved, and that peer review might grind to a halt
- Scientists have a positive attitude towards innovative publication strategies of research data and welcome greater clarity regarding the re-use of their data
- Scientists are sceptical about obligatory measures of data management, since they fear bureaucracy

"Has it been changed in some way?"

Authenticity

Authenticity can be defined[10] as the degree to which a person or system regards an object as what it is purported to be. Authenticity is judged on the basis of evidence. Two issues are presented that support the authenticity of digital data: provenance and persistent identification.

APARSEN has investigated how best to capture and evaluate evidence about authenticity and provenance in a common way that allows the interoperability required to support changes in data holders and processing. A model is proposed, and has been tested, for managing authenticity and provenance throughout the digital resource lifecycle.

APARSEN has analysed results about authenticity and the chain of custody from projects including InterPARES and CASPAR. We have also examined standards relevant for digital preservation, legal requirements and audit and certification of repositories, particularly the revised Reference Model for an Open Archival Information System (OAIS, ISO 14721), ISO 15489, MOREQ and ISO 16363. The practices of a number of repositories have also been examined.

There is clearly a need for more concrete guidelines about what evidence to capture, as provenance, over the whole lifecycle as it is entrusted to a succession of individuals and systems, and perhaps transformed from one format to another. However, the way in which it is captured will undoubtedly change over time, therefore we need a way to combine the various ways used to record this provenance. Since large complex automated systems are impossible to deal with manually, we also need automated ways to deal intelligently with the logical implications of the linkage between the evidence. We also need to be sure that the automated logs about system activities can be dealt with securely.

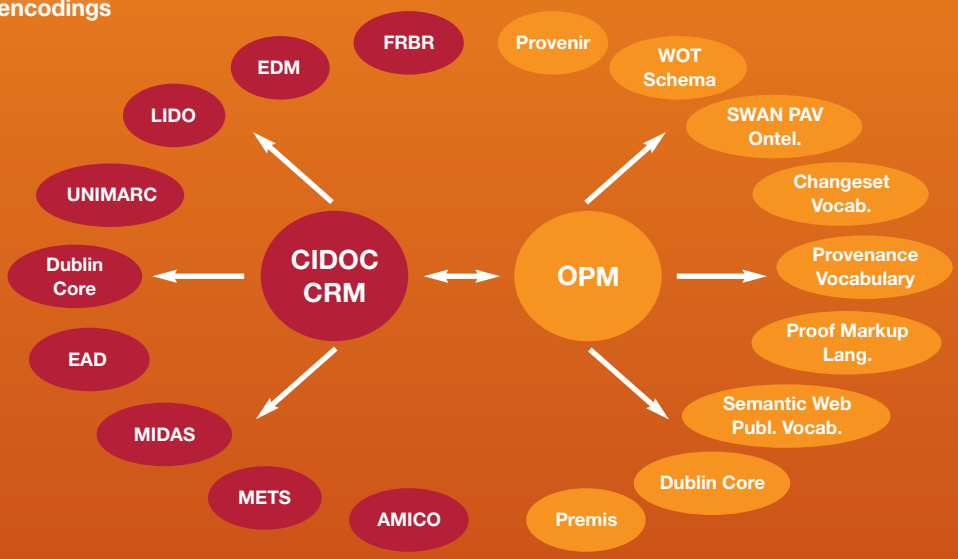
Each of these is addressed in detail in the full APARSEN report[11] and results of the tests are also available.[12]

What evidence?

As for what provenance to capture, in the original definition given in CASPAR[13], Authenticity Protocols (APs) are the procedures to be followed in order to assess the authenticity of specific types of Digital Resource.

APARSEN has complemented this with a set of operational guidelines for an Authenticity Management Policy, i.e. to identify the relevant transformations in the lifecycle and to specify which controls should be performed and which authenticity evidence should be collected in connection with these transformations.

Figure 3
Mapping between Provenance encodings



Evidence must be carefully selected and securely collected, then combined and processed intelligently. How?

The case studies have proved the validity of this approach. On the one hand they have proved to be easily applied and well understood in all the test cases, and on the other hand the simple and yet rigorous concepts introduced by the model may provide a common ground for the management of authenticity evidence and for exchanging it among different systems.

Dealing intelligently with evidence

Provenance interoperability and reasoning is illustrated here, showing how interoperability between the most common provenance capture systems may be achieved. The key contribution of APARSEN is to provide a mapping between the OPM (Open Provenance model) promoted by W3C and CRMdig, an extension of the CIDOC CRM ontology for capturing digital resources – see Figure 3.

As a simple illustration of the sort of automated reasoning that can be carried out, we can for example infer the presence of information in hieroglyphics at a particular event (say a 3D reconstruction) because the part of the column of Ramesses II (that carries said hieroglyphics) was also present at that event – see Figure 4.

This is relatively easy with a small number of facts – but we need the automated rules, which are described in the APARSEN report, to deal with thousands. The rules also allow us to deal with the case that new facts are discovered - suppose that we believed that *George* wrote a document, and therefore wrote all the paragraphs in that document but then we find that *John* instead of *George* is responsible for *writing 1st parag*, we can see that *George* should be replaced by *John* – again, easy for a few instances but difficult for thousands of interlinked facts – see Figure 5.

Secure Evidence

A special security model of provenance data is needed[14]. It contains who contributed what kind of data at a given time and how all electronic resources in an archive are related with each other. This information is needed to prove the custody of some document or file, and to demonstrate what processes interfered with the resource. Provenance data can also be used in order to control the quality of data and it can be used to differentiate original documents from copies.

Four fundamental security requirements have to be fulfilled: confidentiality, authentication, non-repudiation and integrity.

Most of the models outlining an architecture for secure provenance records include the following fundamental properties:

- Provenance data itself has to be encrypted to prevent eavesdropping
- Integrity of provenance data is protected by signatures
- The provenance graph in its complete form is protected by a signature

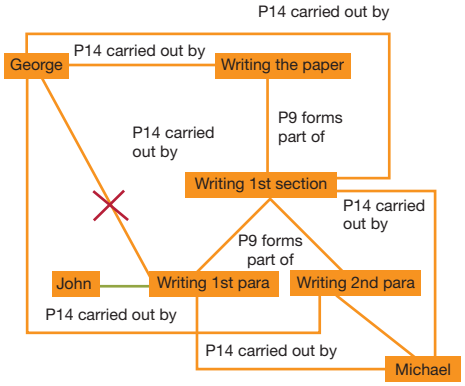
These three criteria have to be fulfilled in order to ensure completeness, validity and integrity as well as confidentiality of provenance data. This sentence should read: 'Although different theoretical models exist, that describe secure mechanisms on how to protect sensitive data, these are often not implemented or used. We describe a number of options but securing sensitive provenance and authenticity still remains a challenging research topic.

Figure 4

Part of a column of Ramesses II

Figure 5

Simple use of inference rules



Key points…

Authenticity is complex. We have provided guidance on:

- What evidence to collect
- How to combine such evidence collected in different systems over time
- How to deal with that evidence intelligently
- How to deal with that evidence securely

"Does the pointer get me to the right object?"

Persistent Identifers and the interoperability challenge

The persistent identification (PI) of digital objects (e.g. articles, datasets, images, streams of data) and non-digital objects (namely real-world entities including authors, institutions but also teams, geographic locations) is becoming a crucial issue for the whole information society. The functionality to unambiguously locate and access digital resources, associate them with the related authors and other relevant entities (e.g. institutions, research groups, projects) is becoming essential to allow the citation, retrieval and preservation of cultural and intellectual resources. The rapid increase of digital assets in recent years, especially in the context of e-science, has made this dependency even stronger, making clear that digital identifiers are crucial for preserving, managing, accessing and re-using huge amounts of data over time.

This is especially true if we consider that today valuable scientific and cultural resources increasingly reside on network-based systems like the Web, encouraging the development of new effective solutions to allow the use of these resources into the future and across many different boundaries (i.e. geographical, organizational, cultural, disciplinary). The implementation of a system for persistent identification of digital and non-digital objects is the first fundamental step to this purpose, becoming a crucial prerequisite for sustained and reliable resource discovery, citation and re-use.

Unfortunately different kinds of identifiers are in use across different stakeholder communities and systems, and multiple identifiers can be available and used within the same system.

On the one hand, it is well-known that the use of URLs (which have been adopted from the birth of the Web to identify and reference network resources) cannot be considered per se a reliable approach to address the long term identification and access of digital resources due to the fact that URLs serve the combined purpose of identifying a resource and describing its location. If the resource is moved to another location, the previous URL is no longer useful to access the resource. For this reason, the use of PIs has become the most popular solution to preserve access to a digital resource regardless of its location, by associating the PI with the correct current location, when the resource is moved.

Some notable solutions for identifying digital objects have been proposed in different domains and several standards are currently at a mature stage of development, like the Uniform Resource Name (URN), the Digital Object Identifier (DOI), the Persistent URL (PURL), the Archival Resource Key (ARK). Unique identifiers for authors are still not commonly used but some author identifier systems and initiatives have started to emerge in the last years, such as AuthorClaim, Scopus Author ID, Researcher ID, arXiv Author ID and ORCID. Recent efforts are also focusing on the development of a standard for uniquely identifying institutions, as addressed by the NISO Institutional Identifier Working Group in the context of information supply chain, but it is worth noting that PI systems for organizations are at a very immature stage.

There are many "persistent" identifier systems. How can they be used together to give us what is needed?

Despite the increasing awareness and interest for PIs, significant weak points still remain, making persistent identification a complex problem, which involves a large number of stakeholders who sometimes have opposing views on many of the issues that need to be addressed. In particular, assurance about the persistence of any of the identifier systems, specifically their resolvers, is lacking.

For instance, user communities such as librarians, archivists, researches, publishers, funding agencies have different visions and approaches to PI concepts, different legal and business models, different requirements and policies. The effect of this differentiation is that some identifier systems turned out to better address the needs of certain communities (and consequently are widely adopted by these communities) but many local solutions are still largely in use to address specific requirements. Thus a discussion on PIs cannot only focus on the technical aspects of assigning PIs to digital resources, but needs to consider the complexity of the entire spectrum of responsibilities and requirements which underlie the development and maintenance of an identifier system. Each of these requirements involves the commitment of many stakeholders to maintain an appropriate infrastructure, to agree on policies, responsibilities, rights and restrictions. Long term funding commitments are, in general, impossible to obtain from funders. This may explain the fragmentation of the current landscape of PI systems and the difficulty of making these identifier systems interoperable.

Since a unique global identification solution is far from being adopted, the challenge is to establish an Interoperability Framework (IF) among the current PI solutions to enable the persistent access, re-use and exchange of information through the use of existing identifiers and associated resources across different systems, locations and services.

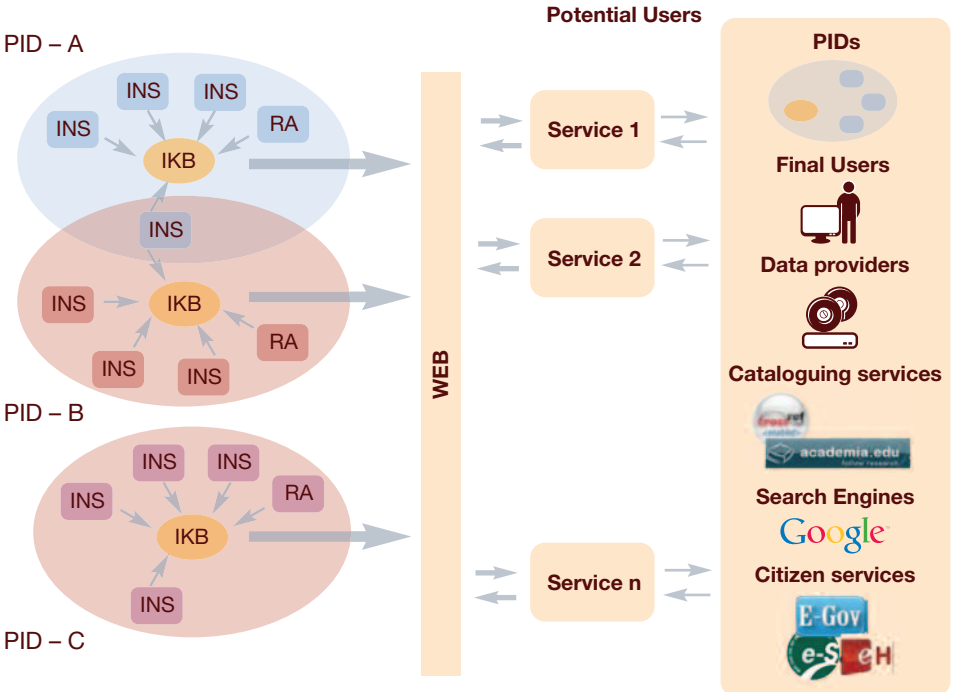
APARSEN has proposed such a general Interoperability Framework (IF) as a starting point to design solutions to support this type of interoperability, and has sought to evaluate this through peer review and practical demonstrations. The next step will be to turn this into a general service through the SCIDIP-ES project (www.scidip-es.eu).

Through the IF – see Figure 6, the identifiers assigned in one context can be encountered, and re-used, in another context, system or time and to access services outside the direct control of the PI assigner. For more information, see the full report.[15]

Key points...

- Multiple independent Persistent Identifier systems are a reality we have to live with since a global system of persistent identifiers is a long way off. We have provided a way to live with this diversity
- APARSEN is creating an Interoperability Framework (IF) for Persistent Identifier systems which enables the persistent access, re-use and exchange of information through the use of existing identifiers and associated resources across different systems, locations and services

Figure 6
The Interoperability Framework for PIs



PID PI Domain *INS Content providers*
RA Registration Agency *IKB Interoperability Knowledge Base*

References

1 See *http://www.trusteddigitalrepository.eu*

2 Data Seal of Approval see *http://www.datasealofapproval.org*

3 See *http://www.nabd.din.de/cmd?level=tpl-art-detailansicht&committeeid=54738855&subcommitteeid=112656173&artid=147058907&bcrumblevel=2&languageid=en*

4 See *http://www.iso16363.org/*

5 Report on Peer Review of Digital Repositories retrieved from *http://www.alliancepermanentaccess.org/wp-content/plugins/download-monitor/download.php?id=Report+on+Peer+Review+of+Digital+Repositories*

6 Morris, C. (Ed.). (1991). Academic Press Dictionary of Science and Technology. London: Academic Press.

7 Refer to Wikipedia article "Standing on the shoulders of giants" retrieved from *http://en.wikipedia.org/wiki/Standing_on_the_shoulders_of_giants*

8 ODE project – see *http://www.ode-project.eu*

9 Report on peer review of research data in scholarly communication, retrieved from *http://www.alliancepermanentaccess.org/wp-content/plugins/download-monitor/download.php?id=Report+on+peer+review+of+research+data+in+scholarly+communication*

10 From the Open Archival Information System (OAIS) Reference Model (ISO 14721:2012), available from *http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=57284* and also from *http://public.ccsds.org/publications/archive/650x0m2.pdf*

11 Report on Authenticity and plan for interoperable authenticity evaluation system, retrieved from *http://www.alliancepermanentaccess.org/wp-content/plugins/downloadmonitor/download.php?id=D24.1+Report+on+Authenticity+and+Plan+for+Interoperable+Authenticity+Evaluation+System*

12 Report on Implementation and testing of an Authenticity protocol, retrieved from *http://www.alliancepermanentaccess.org/wp-content/plugins/download-monitor/download.php?id=D24.2+Implementation+and+testing+of+an+Authenticity+Protocol*

13 See *http://www.alliancepermanentaccess.org/index.php/community/current-projects/caspar*

14 Braun, U., Shinnar, A., Seltzer, M.: Securing provenance. In: Proc. of the 3rd USENIX Workshop on Hot topics in Security (HotSec) (2008)

15 Persistent Identifiers Interoperability Framework *http://www.alliancepermanentaccess.org/wp-content/plugins/download-monitor/download.php?id=D22.1+Persistent+Identifiers+Interoperability+Framework*

APA RSEN

Alliance Permanent Access to the Records of Science in Europe Network

www.aparsen.eu

APARSEN brings together expertise from across Europe including partners from industry, cultural heritage organizations, research bodies and membership organizations and will bring coherence, cohesion and continuity to research into barriers to the long term accessibility and usability of digital information and data. It will defragment ideas about, and create a Virtual Centre of Excellence for, digital preservation.

Airbus Operations SAS *France*

Alliance Permanent Access (APA) *Netherlands*

Austrian National Library (ONB) *Austria*

Centre Informatique National de L'Enseignement Supérieur (CINES) *France*

Consorzio Interuniversitario Nazionale per l'Informatica (CINI) *Italy*

Data Archiving and Networked Services (KNAW-DANS) *Netherlands*

Deutsche Nationalbibliothek (DNB) *Germany*

Digital Preservation Coalition (DPC) *UK*

Estudios y Estrategias, S.A (INMARK) *Spain*

European Organisation for Nuclear Research (CERN) *Switzerland*

European Space Agency (ESA) *France*

Fondazione Rinascimento Digitale (FRD) *Italy*

Forschungsinstitut für Telekommunikation (FTK) *Germany*

Foundation for Research and Technology-Hellas (FORTH) *Greece*

Globale Informationstechnik GmbH (GLOBIT) *Germany*

Helmholtz Association (HA, represented by Alfred Wegener Institute) (AFPUM) *Germany*

IBM Israel, Science and Technology Ltd *Israel*

InConTec GmbH (ICT) *Germany*

International Association of Scientific, Technical and Medical Publishers (STM) *Netherlands*

Koninklijke Bibliotheek (KB) *Netherlands*

Luleå University of Technology (LTU) *Sweden*

Science and Technology Facilities Council (STFC) *UK*

Secure Business Austria (SBA) *Austria*

Space Research Institute of the Russian Academy of Sciences (IKI RAN) *Russia*

Tessella *UK*

The British Library (BL) *UK*

The Stichting LIBER Foundation *Netherlands*

Tieteen Tietotekniikan Keskus Oy (CSC) *Finland*

University of Essex, UK Data Archive *UK*

University of Patras (UPAT), Library & Information Center (LIC) *Greece*

University of Trento (UNITN) *Italy*

APARSEN

Alliance Permanent Access to the Records of Science in Europe Network

www.aparsen.eu

Administrative Coordinator:
Simon Lambert
simon.lambert@stfc.ac.uk

Technical Coordinator:
David Giaretta
director@alliancepermanentaccess.org

Join us on

http://www.linkedin.com/groups/APARSEN-3764755

https://www.facebook.com/APARSEN

https://twitter.com/#!/APARSENproject

APARSEN is co-funded by the European Union under FP7-ICT-2009-6 agreement 269977

SEVENTH FRAMEWORK PROGRAMME

APA ALLIANCE FOR PERMANENT ACCESS

UK Office

Dr David Giaretta
Alliance for Permanent Access
2 High Street
Yetminster
Dorset DT9 6LF, UK

+44 1935 872660
director@alliancepermanentaccess.org

Registered office

Alliance for Permanent Access
Prins Willem-Alexanderhof 5,
2595 BE
The Hague
The Netherlands

Front cover image: European Synchrotron Radiation Facility (ESRF) at Grenoble. Courtesy of Peter Ginter, ESRF.