CASE STATEMENT PROPOSAL RDA WORKING GROUP ON DATA CITATION

v1.5, 3.1.2014

1 Contents

2 Inhalt

3 WG Charter

3.1 Stakeholders

The case statement outlines our work and provides the focus and the boundaries where our research will go. We need to integrate all stakeholders and reflect their views accordingly. So far we identified four stakeholders that will actually use our contributions:

- Data providers – data will be reused
- Solution providers – machine readable data citations
- Researchers – receives citable results
- Community – gains trust and transparency

The beneficiaries will be able to reuse data, reproduce experiments, provide machine readable and machine actionable data citations for complex data sets and trace their data and its usage.

3.2 Mission

Being able to reliably and efficiently cite entire or subsets of data in large and dynamically growing or changing datasets constitutes a significant challenge for a range of research domains. Such data is, for example, generated within research infrastructures during long lasting experiments such as satellite missions, environmental monitoring campaigns, or in permanent installations such as natural hazard detection and early warning systems (e.g., seismic traces acquired by field stations). Several approaches for assigning PIDs to support data citation at different levels in the process have been proposed. These may range from individual PIDs being assigned to individual data elements to PIDs assigned to queries executed on time-stamped and versioned databases or data sets in general. Examples of different strategies in PID assignment currently discussed within research infrastructures are available in [2, 5, 6]

Based on the discussions at the First Plenary Meeting in Gothenburg, the formation of a Working Group on Data Citation (WG-DC) was initiated. The RDA Working Group on Data Citation (WG-DC) aims to bring together a group of experts to discuss the issues, requirements, advantages and shortcomings of existing approaches for efficiently citing

subsets of data. The WG-DC provides both overall minimum requirements for data subset citation, and specifically focuses on a narrow field where we can contribute significantly and provide prototypes and reference implementations. So far different data citation initiatives exist, all of which have their advantages and special purposes. An overview of these standards and their best practices was published by the CODATA Task Group on Digital Data Curation [1]. We encourage strong cooperation with existing initiatives is required: CODATA, OpenAire, DataCite, W3C, Open Annotation Coalition and the related standards.

Our concept includes machine actionable data citations that are efficient and can be applied transparently. For the focused use cases, we will be looking at different types of data and database management systems e.g. SQL-style databases, XML databases / semi-structured databases, graph-based databases, …)

CSV and other static data files (While these are not exactly prominent in settings with large-scale volumes of dynamic data, they are a widely used form of data in many disciplines. Thus, the principles should apply and be transferable as well)

The goal is to assure that each state and subset of data can be uniquely identified in the face of data being added, deleted or otherwise modified in a database, across longer periods of time, even when data is being migrated from one DMS to another. We want to discuss and evaluate different existing approaches to this challenge, evaluate their advantages and shortcomings and identify obstacles to their deployment in different settings, as well as concrete recommendations for the deployment of prototypes within existing data centers. Amongst others these should subsequently form a solid basis for citing data, linking to it from publications in an actionable manner.

Dynamic data citation tackles challenges of versioning and the proper definition of subsets of data in different domains. Potential issues concern the relations between data sets, which need to be captured as well. Other challenges are scalability, costs and benefits (trade off) of ownership and operations that are potentially not reversible.

This WG concentrates on the technical aspects of data citation solutions, i.techniques to make data scalably citeable in a machine-actionable manner using existing PID solutions. It does not address socio-economic aspects of data citation, metadata information to be attached to a data citation, as well as incentives for citing data. We are focusing on proof of concept and prototype implementations. It will collaborate with other RGA working groups on PIDs and other topics under the umbrella of the Interest Group on Data Publication and the PID Interest Group.

3.3 Approach

Currently discussed principles include the following aspects:

- Ensuring that data items added to a data collection are added in a manner that is time-stamped
- Ensuring that the data collection is versioned, i.e. changes/deletions to the data are marked as changed with validity timestamps
- PIDs are assigned to the query/expression identifying a certain subset of the data that one wishes to cite, with the query being time-stamped as well
- Hash keys may be computed for the selection result to allow subsequent verification of identity of the results returned

- Issues such as unique sorting of results need to be considered when the operation returns data as sets and subsequent process work on the sequence the data is provided in

These should be working across all settings where we have a combination of data sources and operations identifying subsets at specific points in time.

We propose a three stage plan consisting of solutions (short-term), plans (mid-term) and the future perspective (long-term).

3.4 Time Frame:

The work plan fort he WG consists of three phases. All of them will be accompanied by networking and intense feedback loops between the WG-DC, the data providers, the solution providers and the community. Fruitful collaboration with other initiatives such as CODATA, DataCite, W3C, open Annotation coalition and others is carried out along all three phases.

3.4.1 Short-term goals (M12)

- Evaluation of recommended data citation approaches for specific scenarios
- Selecting pilot candidates in cooperation with stakeholders/data owners
- Detailed planning of technical aspects of data citation approach preparing for implementation
- Develop a model for long term proof data citation within databases and initial prototypes and demonstrators.

3.4.2 Mid-term goals (M18)

- Develop a set of reference implementations of the data citation model for selected pilot data types.
- Evaluate developed prototypes
- Establish consensus on a universal data citation model that can be implemented independently from specific systems or vendors.

3.4.3 Long-term goals (> M18)

- Seek official endorsement for data citation within the RDA community and foster the application of identified standards.
- Become the contact point for data citation questions and provide the community with best practices and know how.

3.5 Pilots

Initial candidate pilots identified so far include

- LNEC Infrastructure Sensor Network Data, Portugal (http://www.lnec.pt/)
- SCOR/IODE/MBLWHOI Library Data Publication Project: Marine Biological Laboratory /Woods Hole Oceanographic Institution (MBLWHOI) Library

(http://darchive.mblwhoilibrary.org/) and the British Oceanographic Data Centre (BODC) (https://www.bodc.ac.uk/data/published_data_library/)
- Million Song Database (MSD), Vienna University of Technology, MIR team
- DKRZ, Germany / Earth System Grid Federation (PCMDI, BADC, DKRZ and others)
- Field Linguistics (The language archive, http://tla.mpi.nl), archive of MPI for psycholinguistics, http://corpus1.mpi.nl

4 Value Proposition and Beneficiaries

4.1 Value Proposition

Digitally driven research is a rather young discipline that evolves fast. As a result the tools and the data are rarely developed with a focus of long term awareness. What matters most to researchers are fast results and prompt publications. Whether the data they produce today can be understood, interpreted or even accessed in the future is receiving too little attention at the moment, as short term results are in the focus. Only if results can be reproduced precisely, the validity of research experiments and business processes can be judged, evaluated and verified in a machine-actionable manner that is scalable and can cope with dynamically changing data. Hence there is a strong need for data citation mechanisms, standards and styles that allow identifying portions of large data set with a precision, regardless of the format of the original dataset or the technical implementation of the data retrieval and storage in the data centre.

An additional challenge within the area of research data is the requirement to cite evolving data reliably. Researchers need the possibility to reference data material that is subject to change. Hence mechanisms are required that allow to cite data as the used it during a particular experiment. When the data gets updated, modified or deleted, these changes must be reflected by the citation as well. Therefore versioning of the data is an important factor. Also, the possibility to specify subsets and derived data is a requirement. Being able to identify, reference, share and distribute specific subsets encourages reuse amongst researchers. The easier and more transparently this citation process can be implemented, the higher is the acceptance among the target audience and the designated community. We will provide proofs of concept, mockups and prototype implementations that can be tested and used by the community. We want to go beyond theoretical work and deliver real world applications for our models. In an optimal setting, a researcher, when selecting a subset of data for an experiment, will be issued with a PID that allows others retrieving the same data set again.

The international orientation and interdisciplinary nature of the RDA community provides input from various interesting areas, enabling broad research and application of the results of the WG-DC. Participation of researchers, data providers, solution providers and the community allow integrating expert knowledge from various perspectives. This direct access to domain experts boosts the development of new standards within the area of data citation and allows improvement via direct feedback loops. Collaboration with other initiatives in the field is also a key concern of this WG.

4.2 Beneficiaries

Individuals, communities, and initiatives that will benefit from the RDA WG on Data Citation.

- Researchers: by being able to cite their data, by being able to reuse data

- Database developers: by reducing redundancy
- Digital preservation managers: by being able to retrieve subsets of data
- Data centers: by enhancing reuse of existing resources
- Data managers and data scientists: by having tools for referencing subsets in dynamic data environments
- Professional societies: by being able to reproduce experiments
- Publishers: by encouraging reuse and an increased level of trust
- Repositories, data archives: by being able to reference, cite and retrieve subsets
- Software tool developers: by allowing transparent implementation of data citation capabilities
- Funding agencies: by promoting better science (by enabling repeatability and verification) and by safeguarding the investment made in data by fostering re-use

4.3 Key impacts of the RDA Data Citation initiative

- Provide the knowhow for data citation of partial datasets from dynamic data sources
- Enhance reproducibility of research results by allowing peers to re-execute experiments
- Facilitate discovery, access, and reference of large data sets
- Provide a reference model for dynamic data citation
- Enhance digital preservation of data sets and their reference

4.4 Engagement with existing work in the area

- CODATA1
- OpenAire2
- DataCite3
- W3C4
- Open Annotation Coalition 5
- FORCE11
- and others

5 Work Plan

5.1 Work plan components

5.1.1 Analysis of requirements and the selection of candidate solutions (m1-6)

In the beginning phase we will consult existing work in the area of data citation and study available best practices [1] and of the methods used already in the existing databases [2, 5, 6]. We describe the minimum requirements for a successful data citation for a dynamic dataset, and specify also additional optional parameters with prioritization for use in such citations. This work is done in close co-operation with research infrastructures, data centres

and publishers, to maintain realistic resource requirements and immediate applicability. We will evaluate recommended data citation approaches for specific scenarios and select pilot candidates. The selection of these candidates will be in close cooperation with stakeholders and data owners to make sure that the developed methods are practical to implement. Detailed planning of the technical aspects of data citation approaches will guide us during the implementation.

5.1.2 Defining reference models (months 4-12)

We will develop a technical reference model for data citation, including case studies for use in relational database systems. The general model should be open, extensible and implementation agnostic, is compatible with existing data citation methods (i.e. DOIs or EPIC), and will retain significant part of the technical implementation open for the data center as long as the minimum requirements are fulfilled. This includes an analysis of appropriate granularities for publishable data set extracts. This will result in recommendations / good practice examples of how to make data versioned and time-stamped in different settings based on existing approaches/solutions.

5.1.3 Improve and test the model iteratively (months 8-15)

The model developed in the previous step will be evaluated against suitable data sets of considerable site from various research disciplines. Similarly, the general minimum model approach is tested using data centres without such relational database structure. The consortium partners will be asked to provide their real world research sets as a testbed upon which the model can be evaluated. This will follow an iterative approach on order to allow improvements.

5.1.4 Promotion of the RDA Data Citation Model and Reference Implementation (months 12-18)

Promotion activities will include wide-spread dissemination about the data citation model. The ready to use reference implementation will be accompanied by substantial documentation and use case scenarios in order to increase acceptance and encourage contributions.

5.2 WG-DC operation

5.2.1 Form and description of final deliverables

D1 Definition of minimal requirements of the data citation for subsets of dynamic data

D2 Model for citing subsets of dynamic data in relational databases

D3 Case study of implementing data citation mechanism different data types

D4 Reference architecture document for citation of dynamic datasets

5.2.2 Milestones

M1 Agreed principles of making dynamic data citable

M2 Recommendations on approach/implementation of dynamic data citation support available

M3 Community awareness and acceptance

6 Initial Membership

6.1 Leadership (brief biographic notes in Appendix A):

- Co-Chair: Andreas Rauber, Vienna University of Technology & SBA
- Co-Chair: Dieter van Uytvanck, MPI
- Co-Chair: Ari Asmi, University of Helsinki

6.2 Members/Interested:

- Daan Broeder, MPI

- Hans Pfeiffenberger, Alfred-Wegner Institute for Polar and Maritime Research

- Peter Wittenburg, Max Plank Institute

- Jeroen Rombouts TU Delft

- Joachim Wambsganss Uni Heidelberg

- Ilya Zaslavsky UCSD

- JuanLe Wang Chinese Academy of Sciences

- Robert H. McDonald Indiana University

- Emily Grumbling NSF

- Diana Hendrickx Maastricht University

- Stefan Pröll SBA Research

- Patricia Cruse DataCite

- Martina Stockhause WDCC/DKRZ

- Christoph Becker TU Wien

- Ari Asmi Uni Helsinki

- Natalia Manola University of Athens

- Constantino Thanos ISTI-CNR

- Volker Boehlke Uni Leipzig

- Thomas Eckart Uni Leipzig

- Paul Uhlir National Academy of Sciences

- Yannis Ioannidis University of Athens

- Shih-Chieh Ilya Li Academia Sinica/CODATA Taipei

- Jane Hunter, University of Queensland

7 References

[1] CODATA Task Group on Digital Data Citation. Best Practices: Research & Analysis Results, CODATA, 2012.

[2] Huber et al., Data citation and digital identifiers for time series data / environmental research infrastructures. Draft document, 2013

[3] S. Pröll (Ed.). Position statements submitted to the BoF Session on Data Citation. at the Research Data Alliance - Launch and First Plenary, Gothenburg, Sweden, March 18-20, 2013. http://forum.rd-alliance.org/download/file.php?id=90

[4] S. Pröll and A. Rauber. Minutes of the BoF-Session on Data Citation, at the Research Data Alliance - Launch and First Plenary, Gothenburg, Sweden, March 18-20, 2013. http://forum.rd-alliance.org/download/file.php?id=110

[5] S. Pröll and A. Rauber. Scalable Data Citation in Dynamic, Large Databases: Model and Reference Implementation. In Proceeding of IEEE Big Data 2013, Santa Clara, CA, 2013.

[6] Puneet Kishor: Serving Data, Licenses, Citations, and Tracking Use, March 2012, http://punkish.org/Serving-Data,-Licenses,-Citations,-and-Tracking-Use

8 Appendix A: CVs

Leadership Biographical Notes

8.1 Andreas Rauber

Andreas Rauber is Associate Professor at the Department of Software Technology and Interactive Systems (ifs) at the Vienna University of Technology (TU-Wien). He furthermore is president of AARIT, the Austrian Association for Research in IT and a Honorary Research Fellow in the Department of Humanities Advanced Technology and Information Institute (HATII), University of Glasgow. He received his MSc and PhD in Computer Science from the Vienna University of Technology in 1997 and 2000, respectively. In 2001 he joined the National Research Council of Italy (CNR) in Pisa as an ERCIM Research Fellow, followed by an ERCIM Research position at the French National Institute for Research in Computer Science and Control (INRIA), at Rocquencourt, France, in 2002. From 2004-2008 he was also head of the iSpaces research group at the eCommerce Competence Center (ec3).

8.2 Dieter van Uytvanck

Dieter Van Uytvanck (1980, Belgium) studied computer science at Ghent University and linguistics at the Radboud University Nijmegen and wrote an MA thesis on analyzing medieval charters with data mining. Based at the Max Planck Institute for Psycholinguistics, he is involved since 2008 in the construction of the CLARIN research infrastructure for Humanities and Social Sciences (www.clarin.eu). As of 2012 he is the CLARIN ERIC director responsible for technology.

8.3 Ari Asmi

Ari Asmi is Research Coordinator at the Department of Physics at the University of Helsinki. He coordinates the data infrastructures of atmospheric and ecosystem research data produced by the Finnish National Centre of Excellence in Atmospheric Science (ATM). He is

responsible for development of data policies, practices and methodologies in the Division of Atmospheric Sciences and is the leader of development for common research strategy of European Environmental Research Infrastructures in the ENVRI EU project. He received his MSc and PhD in Atmospheric Physics from the University of Helsinki. He is a board member of the Finnish Association of Aerosol Research.