Dynamic Data Citation

Making Evolving Research Data Citable

Stefan Pröll
Secure Business Austria
Vienna, Austria
sproell@sba-research.org

ABSTRACT

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. Current approaches rely on pointers to entire data collections or on explicit copies of data. They do not scale with large quantities of data. Hence a new method is required that enables to create, reference and site arbitrary subsets of research datasets in a scalable way. Developing such a method and providing a framework for its application within multiple research data scenarios is the goal of this thesis.

Categories and Subject Descriptors

H.3.7 [**Information Systems**]: Digital Libraries—*Collection, Dissemination, Standards*; H.2.8 [**Information Systems**]: Database Management—*Scientific databases*; H.2.0 [**Information Systems**]: Database Management—*Security, integrity, protection*

Keywords

Research data management, Data citation, Persistent identification, Relational Database management systems

1. CONTEXT AND MOTIVATION

Modern research is data driven and the amounts and complexity of the data produced in various disciplines is increasing [4]. Research data is not just a by product of the paper publication process anymore, it has become an important academic asset in its own right. Not only are large scale research facilities such as the CERN producing petabytes of data every year[1], also researchers with commodity hardware work with increasingly large amounts of heterogeneous data sources.

We need to share this data in order to advance research further. This goal can only be achieved if we can identify, reference, cite, reuse, verify and validate research data. Therefore mechanisms that allow to treat data as academic output as first-class objects are required [1]. Data needs to be treated with the same devotion and detail as it was granted and established many decades ago for traditional paper publications. The view that I have on data citation in the context of my thesis must not be confused with questions dealing with citation styles and citation formats. The goal of my thesis is enabling researchers to precisely identify and share specific datasets in a lightweight fashion in order to enable reproducible research. What we need is a flexible approach, which supports scientists during their work with data in highly dynamic environments.

In many cases research data is abandoned after the paper publication process has been finished. If the data is considered at all, then in most cases the researchers put their data in one large zip archive file and store it at an institutional Web server. The archive gets referenced with a standard URL and the researcher moves on with his work. In recent years, the topic of data citation has received more attention [10] and the management of data is getting professionalised. Many approaches are now using persistent identifiers that allow the identification of entire datasets, but still this solutions are based upon indivisible datasets that are potentially huge in size and do not provide flexibility for other researchers to reuse the data efficiently.

PID approaches are suited very well for static data, which should only serve as reference point once it has been created. Using the identification and additional metadata is sufficient to search, identify, and retrieve data again. PID systems usually facilitate delegating methods that outsource the resolution of the actual data. The metadata and the links to the data are stored and maintained centrally in many cases [12]. The relationship between metadata and source data requires to be managed over the long term. The persistent identifier never changes by definition, only the location of the digital object may be updated. So far, PID systems do not provide any interfaces for data interaction, their focus is on resolving locations. Although some PID systems (e.g. ARK [7]) provide simple means for referencing sub sets, there are no mechanisms available that allow the definition of subsets where the data is allows to evolve. Currently sub sets are often mere a hint provided in textual form, rather than a machine actionable data structure that could be directly reused. Assigning persistent identifiers (PIDs) to data portions of finer granularity, i.e. database rows or even cells

[1] http://home.web.cern.ch/about/computing

would require enormous numbers of unique identifiers and yield unfeasible citations.

However, many settings require us to go beyond these limitations and introduce scalable and machine-actionable methods that can be used in dynamically changing, very large databases. Also many datasets continue to grow and are updated as the datasets are used productively in experiments. In order to enable data citation in dynamic environments versioning support is required. Furthermore different stakeholders may be interested in diverse portions of the data. Hence, clearly defined subsets of the data need to be identifiable and citable as well.

The goal of this thesis is to provide a generalised model for dynamic data citation and several reference implementations demonstrating the feasibility of the model. The purpose of my work is to enable data creators, data centres and data publishers with the possibility of referencing all their data assets in arbitrary detail for the long term.

2. BACKGROUND RELATED WORK

Publications increasingly contain references to data that was used or generated during the research that substantiates the work. However, research datasets are often treated as one entity, i.e. indivisible, static and referencable as one unit. In many cases, data is referenced bibliographically, the data itself is then often deposited at an institutional site and referenced by providing an URL. Obviously this mechanism is not suitable for sustainable data citation for several reasons. Uniform Resource Locators (URL)[2] have not been designed to be stable for the long term. As their name implies, URLs refer to a location, not the object itself. As a result, many URLs that served as data citation reference are not accessible any more. Either because the author of the dataset left the institution and the Web page was taken down, or because the server moved and the location changed.

To overcome the problem of changing locations, the concept of persistent identifiers was introduced. Persistent identifiers (PIDs) provide unique identification of digital objects and reliable locations of Internet resources. PIDs require organisational effort for the management for the linking between the data and the identifier. Also, services for locating and accessing objects are necessary. The organisations providing these services are denoted Registration Authorities (RA). These RAs are responsible for the long term access, resolution and maintenance of the identifiers they issued for digital objects. There exist different solutions for the implementation of persistent identifiers. The authors of [3] provide an overview of the most common approaches.

Although persistent identifiers solve the problem with locations of digital objects, there are drawbacks for dynamic data: Subsets require their own identification and metadata. Assigning PIDs to data portions of finer granularity, i.e. database rows or even cells would require enormous numbers of unique identifiers and yield unfeasible citations. Also PID approaches are suited very well for static data, which should only serve as reference point once it has been created. However, many settings require us to go beyond these limitations and introduce scalable and machine-actionable methods that can be used in dynamically changing, very large datasets. Also, many datasets continue to grow and are updated as the datasets are used in experiments. In order to enable data citation in dynamic environments versioning support is required. Furthermore, different stakeholders may be interested in diverse portions of the data. Hence, clearly defined subsets of the data need to be identifiable and citable as well.

3. PROBLEM STATEMENT

The approaches solely based on persistent identifiers can not be applied on evolving data. The requirements for dynamic data citation I am interested in, are driven by the following research questions:

- What are the fundamental properties that render any subsets of dynamically changing research data citable for the long term?
- Does there exist a general model that can be applied to various data formats, data stores and data types?
- How can dynamic data citation scale up with growing amounts of data?
- What semantics are required for enabling machine actionable data citation?
- How can data citation of subsets be made transparent so that users are not hindered by citation mechanisms?

These research questions are always complemented by the requirements that the community has for their data to be citable. Although the focus is clearly on a technical solution, the community acceptance is also highly relevant for all potential solutions.

4. RESEARCH GOALS

My engagement in the area of research data was started while working at the APARSEN[3] project. During this project I was able to contribute to work packages dealing with authenticity and provenance [17], peer review of research data [11] and data annotation, reputation and quality [13]. Based on the knowledge I could gain during the work on APARSEN my goal was to apply certain aspects of research data in a more dynamic environment. The TIMBUS[4] project provides such a setting, where we could analyse workflow management systems for their ability to produce and maintain the provenance data [8]. Workflows are widely used in science and business to express a sequence of steps that needs to be processed in order to a achieve a predefined goal. Provenance data helps to trace and understand how data was produced and transformed during such a workflow. Hence it constitutes to validation and verification, another topic that we are working on within the TIMBUS project [19]. Therefore it is important for preserving not only data but processes with their dynamic nature and enhance their reproducibility and their re-usability.

[2] www.ietf.org/rfc/rfc1738.txt

[3] www.alliancepermanentaccess.org

[4] www.timbusproject.net

The link between the topics I could gain insights within APARSEN and TIMBUS then quickly became obvious: There exists a gap between static data citations as they are currently used in scholarly communications and the requirements of highly dynamic data as it is actually used today. Many different areas of academia, science, business and public life already use highly dynamic data assets. Emerging areas Big Data or data-driven journalism analyse, curate and share increasing amounts of data that can not be processes with traditional methods anymore. This entails that classical referencing customs will not be able to deal with the new data paradigm that has already fully arrived. Citation methods need to adapt to the new challenges that we are currently approaching. We need concepts, methods and tools that allow us to re-use, re-analyse and re-produce dynamic data. Static data references or predefined subsets are not suitable anymore. What is needed is a more flexible approach, not only allowing to use, but also fully integrating the dynamic nature of data and utilising the opportunities that dynamics offer. We need to detect, maintain and preserve high quality data, that is produced in enormous quantities. Therefore we require citation systems that allow us and future researchers the reliable reference, search and retrieval of the data assets we produce today for their future analysis, interpretation and re-interpretation.

The goal of my thesis is to contribute to a new data citation paradigm that allows to reference and therefore reuse, reproduce, validate and share dynamic data assets. The first step towards this goal have already been achieved. These are described in Section 5. There are many areas which would benefit from a dynamic data citation mechanism. It is not only the scholarly publishers and researchers that require precise references. As our society is relying more and more on digital data, we need means to find specific portions of this data again. To reuse it in the future and to validate its use in the past. Therefore what I aim to achieve is to provide a framework for dynamic data citation, which allows various stakeholders to apply easy to use but still highly precise data citation to their data. Citing data needs to be automated, read- and interpretable by machines as well as by human beings. Ambiguity needs to be vanished such that data becomes accessible for the long term. The benefit of a uniform, generalized data citation model is secure, easy and persistent access to data.

5. PLANNED APPROACH AND CURRENT STATUS

The research for my thesis has already passed the initial phases, where I tried to find and define the topic of my work. These phases are described in the following sections. Future work is outline in Section 5.5 and Section 5.6.

5.1 A Basic Model for Citing Datasets in Relational Databases

We introduced a case statement highlighting the fundamental ideas at the Research Data Alliance Plenary Meeting [14] in Gothenburg in February 2013 and aim to establish a working group under the umbrella of the RDA. We collected contributions from related research areas and institutes, in order to get an overview of current challenges in other areas producing high quality scientific data. Thereby we hope to motivate others to contribute and share their experience and provide us with further use cases we can investigate in.

We began to develop a basic model focusing on relational database management systems (RDBMS) which support many of our requirements off the shelf. These databases can be used to retrieve arbitrary subsets of data. Hence we concentrated on this database model for a first pilot study before discussing the general applicability. Our model is based on timestamped SELECT-Statements and versioned data. Queries can be used in order to persistently identify subsets of arbitrary complexity and size. The dynamic nature of research databases requires mechanisms that allow to trace and monitor all changes that occurred during time. Hence, temporal aspects have to be included in the model. This timing information needs to be stored on each UPDATE, INSERT or DELETE statement for the affected records, enabling to trace all changes that occurred. As relational database systems are set based, sorting is not an inherent criteria automatically. Therefore, we need to specify stable sorting criteria that are automatically applied to the subsets. Depending on the size of the dataset, the schema and the complexity of the query, the retrieval of the result set can challenging. If these properties are met, citing only the query persistently is sufficient to meet our requirements. It guarantees not only consistent result sets across time, but also consistent result lists even in case of none or ambiguous result set sorting in the initial query, even in the case of migration to a different DBMS.

5.2 A Position Statement for Citing Datasets in Relational Databases

In [15] we provided a position statement describing our refined model for citing datasets by focusing on RDBMS. We again concentrated on the queries and their results, not on the large, indivisible data portions as a basis for reference. In this work our model provides guidance on how to enhance the data model used for processing research data, in order to ensure it can be reliably cited and re-used in the future.

In temporal RDBMS, timestamps are provided for all records. This ensures that specific versions of data can be retrieved without having to stall the database tables for additional data. As records can change, they need to be versioned, i.e. all changes that affect the data need to be traceable. This entails that statements such as DELETE or UPDATE must not to destroy the data, but rather set markers that indicate that a record has been marked for deletion or that it as has been updated by a more recent version.

The construction of subsets of complex databases can be easily be achieved by issuing SELECT-queries against the RDBMS. To enable the data citation facilities, the SQL-query has also to be augmented with a timestamp. This timestamp maps the subset to a specific state of the data. As the records in the database can be altered individually, it needs to be ensured that the correct version that was valid at the query's timestamp is selected for inclusion in the subset. Hence the timestamp of the query can be used to retrieve arbitrary subsets of a specific version of the data.

There are several possibilities how this version information can be implemented [18]. The temporal timestamp contains the explicit date at which the data has been changed. Suitable timestamps are dates that are granular enough to capture the point in time that enables to differentiate between two versions of data. The actual chronon to be picked depends on the potential frequency of changes in data, which is not a trivial task [5]. Thus granularity can range from days to milliseconds. Snodgrass et al. differentiate between valid time and transaction time [6]. Valid time refers to the period until the data was considered a true fact in the database. Transaction time refers to the time when the change occurred on the system, independent of its temporal meaning for the actual data. The valid time concept is a reference to the real world, the transaction time only refers to the system time, at which a change of data was manifested.

It is essential being able to identify all records uniquely. This property can also be handled by any RDBMS with the concept of primary keys. Hence our citable database schema requires each table to be equipped with a primary key. Primary keys are by definition unique, hence it allows to specify a unique sorting of the records to be included into the subset. Otherwise the sorting may depend on the features or settings of the database management system. Thus the sequence in which the results are delivered could change and the result set could be interpreted differently by post-processing steps. Therefore we need to establish stable sorting by specifying a standard sorting order based on the primary keys on each query.

These queries themselves need to be stored and augmented with a timestamp that reflects the time when the query was issued. The query's timestamp defines what versions of the records are included in the subset. A hash function over the SELECT-Query allows to identify queries that have already been issued against the system. Then a mechanism to identify the queries and the subsets they produced is required. In this case, PIDs become very useful, as a query that identifies a precise subset is static. If no changes of the records have occurred between two runs of the identical query, the same PID needs to be assigned to both runs of the query. It is easily possible to automate the creation of timestamps for data altering events as well as for queries. This allows to implement dynamic data citation transparently, i.e. no specific action is required on the user side: whenever a researcher selects subset of data for an experiment, the data is returned with a PID. This ensures that upon re-invoking the query, the PID is identified and an identical set is returned, even in identical order.

The model we introduced in this section describes how arbitrary subsets of data in potentially large relational databases can be created and retrieved at a later point of time. We then investigated in a more generic model that can address data sources of various types.

5.3 Generalisation: Expanding the Model to Data Sources

The model that enables dynamic data citation introduced in Section 5.2 is not limited to relational databases. As nicely generalised from [14] in [9], the core concepts themselves can be mapped to other data models as well. The following requirements enable dynamic data citation on a generic level: The basic requirements are uniquely identifiable data records, that can be included in subsets of data. These records that form a subset need to be identifiable on an individual level. Furthermore, a versioning scheme must be available. These versions should reflect events such as insertion, updates or deletion. Hence no change on the data must be lost, regardless what data model is used. The versioning mechanism should include timestamps that allow to derive the set of valid records at a given point of time. For constructing subsets, the data source must provide a query language, which is powerful enough to select specified records based on precise criteria. To enable citation of subsets, it is sufficient to store the queries that led to the subsets and combine them with the timestamp. This timestamp provides the mapping between the query and the different versions of the records. This query is the key to the subset. Hence the query needs to be identifiable in order to retrieve the subset at a later point in time. With the requirements introduced, arbitrary data sources can be cited. The model based on these requirements allows to cite data that is evolving within the data source.

Besides these criteria introduced, there are additional considerations that have to be made. The requirements mentioned so far only consider internal properties of the system the data resides in. It is clear that external influences that can alter data, but are not recognised by the data storage system, need to be prevented. Furthermore, side effects that depend on the query system, the query language or specific properties of the datasets need to be removed in order to enable reproducibility. If the query language provides functions that are based on non-deterministic calculations, they have to be treated. This includes all sorts of randomised functions (e.g. a random number generator) or relative time specifications (e.g. CURDATE()). Such operations hinder the re-execution of a query for retrieving the exact same result, as they depend on external influences. How this issue can be mitigated will be part of our future work. Schema or format changes are a challenge that needs to be addressed.

5.4 Application on a Use Case

In [16], we derived a use case for dynamic data citation from from the Portuguese national civil engineering laboratory (LNEC) that is currently one of our partners in the TIMBUS project. In this use case queries are used for selecting the data that is the basis for the production of a monitoring report. According to our model these SELECT-statements can be used for data citation as long as some additional properties are considered.

As data usually evolves over time it is important that the data citation model supports basic DML statements that allow to manipulate existing data. Still it needs to ensure that citations are possible, that only include the records within their appropriate state that has been valid at the time of the original query [2]. All data manipulation language (DML) statements such as INSERT, UPDATE or DELETE are provided with the timestamp of their execution. Data is never deleted (except in the case of e.g. legal requirements for deletion, which need to be recorded as invalidating the authenticity of affected subsets of the data) and the previous state of a row within a dataset is maintained during the whole life cycle. Hence the state of the dataset that was valid at the execution time can be reconstructed. As a result, the evolution of data is traceable, citations can be managed in a scalable and automated way and the model does not require vendor specific features.

The generation of citable result sets is query centric, which entails that not the data itself needs to be explicitly stored, but the query statements that define the elements of a subset are kept. Our model supports the citation of both static and dynamic data that gets updated and changed. This provides a very flexible citation solution that enables the provision of references in a timely manner. Identity of result tuple sequences is ensured by enforcing predefined sorting criteria for queries not explicitly defining unique sorting themselves. Thus result sets are stable across time and consistent, even in case of lacking or ambiguous result set sortings in the initial query or in the case of changes in the internals of the database management system.

Assigning PIDs to the query allows data citation of arbitrary subsets of very large datasets. Our model is agnostic of the actual PID-system used because only plain text queries will be referenced. The query can then be used to retrieve the corresponding subset of data in the correct version. The time stamp annotated to the query allows determining which version to choose. Such a query can then be rewritten in order to retrieve the appropriate state of the database at the desired time frame.

A further detail that is crucial for the acceptance of dynamic data citation is transparency and the effort required for implementing the model for a given data source, both from the point of view of the operator as well as from a user's perspective. We proposed three different approaches for storing the required metadata, all described in a bit more detail, comparing their advantages and disadvantages. The implementation of this model should be as non-invasive with regard to the existing applications interacting with the DB.

5.5 Next Steps: Broaden the Approach and Engage the Community

The concepts are currently considered to be addressed as part of a larger working group within the Research Data Alliance (RDA[5]). Our goal is to provide proofs of concept, mock-ups and prototype implementations, that can be tested and used by the community within the near future. A first prototype will be implemented by inserting the query rewriting and time-stamped storage. This allows users of relational databases to create citable subsets and share them with peers.

My future work will focus on on other data formats that are widely used within research. This includes specialised file formats from various disciplines and areas, such as XML data structures, flat files but also various database models from the NoSQL domains that are gaining more and more attention recently. Examples for such database models are key-value stores, graph databases, document databases or various tuple stores. These systems are often deployed in highly specific environments. My plan is to apply the generic dynamic data citation model to such systems and test its feasibility for reliable and persistent data citation.

The success of new data citation models relies upon the acceptance of the community. Only if the approach developed during my work does actively help and support researchers, scientists, publishers and data centres for citing their data assets, it will be accepted. Hence close collaboration with different stakeholders is crucial in order to understand their needs and provide solutions that sustain.

5.6 Time frame: Whats Happening Next?

Working on this thesis is the largest project I encountered so far. For the sake of orientation and planing, I structured the work for my thesis into short-term and long term goals. The short-term goals serve as a basis for the work I am planning within the horizon of my thesis.

Short-Term Goals. include the evaluation of recommended data citation approaches for specific scenarios. Therefore I want to analyse the potential stakeholder that require data citation mechanisms. This research will also include research data types, storage and usage. Also I will investigate how the stakeholders currently reference their data and analyse which PIDs are most widely used and if advanced technologies such as machine readable and actionable identifiers are available. A very important questions regards the stakeholder's requirements, needs and wishes towards their research data and its citation.

This phase will be followed by the selecting pilot candidates in cooperation with the aforementioned stakeholders and data owners. Then, detailed planning of the technical aspects of data citation approaches are required. These should be considered for implementation tests. Then I will develop a model for long term proof data citation within databases and initial prototypes and demonstrators. This is the phase I am currently working on.

Mid-Term and Long-Term Goals. are based on the requirements I could gather so far. I want to develop a set of reference implementations of the data citation model for selected pilot data types. Then these developed prototypes will be evaluated, preferably in close cooperation with the stakeholders. As a goal of this phase I aim to establish consensus on a universal data citation model that can be implemented independently from specific systems or vendors. There are many practical applications that require precise and flexible data citation. The range of stakeholders reaches from publishers who want to augment and enhance publications with actionable datasets to businesses who require to identify which data was used during processes. The framework developed during the work on this thesis needs to be evaluated against all these scenarios and improved iteratively. The frameworks will be validated against the use cases provided by the community.

6. EXPECTED RESULTS

The expected results are derived from the research questions given in Section 3. I will tackle these questions with

[5] www.rd-alliance.org/working-groups/data-citation-wg.html

the approach outlined in Section 5 by using the road map outline in Section 5.6. As intermediate result I expect to have a thorough analysis of the stakeholder of data citation, their needs, requirements and expectations. This survey will provide the insights that are required to analyse the use cases that would ideally also be distilled from the stakeholder analysis. As outcome of this step I want to provide several scenarios from different areas that require dynamic data citation. I can imagine that many of the use cases require highly individual solutions. When working on these, the generic model for data citation I envision will be refined further until the essence of data citation in evolutionary data environments becomes clear.

The insights gained during this process will then be used in order to create with the collaboration of the use case owners a set of best practices that can be applied only with minor modifications within each scenario. The implementations could then be refined and iteratively be improved in order to serve with stable, secure and reliable data citation facilities for various applications.

7. CONCLUSION

Digitally driven research is a rather young discipline that evolves fast, what matters most to researchers is fast results and prompt publications. As a result there is a lack of long term awareness. Currently data is treated rather as supplement to publications than as a first class research object. The research data which is produced today needs to be understandable, interpretable and accessible in the future. We want to support this paradigm and highlight the need for long term proof data citation capable to be used within highly dynamic research settings.

Hence we are working on a model that is generic but yet specific enough to be applied to various data scenarios. We began our research focusing on relational database management systems and extend this framework for supporting various data sources that are used within today's research. We plan to identify requirements that enable data sources to provide citable subsets of data even if the data is still evolving. We envision automated, highly transparent and pragmatic data citation that keeps research data accessible for the long term.

8. REFERENCES

[1] A. Ball and M. Duke. Data citation and linking, 2012. Digital Curation Centre.

[2] P. Dadam, V. Lum, and H. Werner. Integration of time versions into a relational database system. In *Proceedings of the 10th International Conference on Very Large Data Bases*, pages 509–522, 1984.

[3] J. K. Hans-Werner Hilse. *Implementing Persistent Identifiers: Overview of concepts, guidelines and recommendations.* Consortium of European Research Libraries, London, 2006.

[4] T. Hey, S. Tansley, and K. Tolle. *The Fourth Paradigm: Data-Intensive Scientific Discovery.* Microsoft Research, 2009.

[5] C. S. Jensen and D. B. Lomet. Transaction timestamping in (temporal) databases. In *Proceedings of the International Conference on Very Large Data Bases*, pages 441–450, 2001.

[6] C. S. Jensen, M. D. Soo, and R. Snodgrass. Unifying temporal data models via a conceptual model. *Information Systems*, 19:513–547, 1993.

[7] J. Kunze. The ARK Identifier Scheme. RFC, Apr 2013.

[8] R. Mayer, S. Proell, and A. Rauber. On the applicability of workflow management systems for the preservation of business processes. In *Proceedings of the 9th International Conference on Digital Preservation (iPres 2012)*, 10 2012.

[9] R. Moore. Workflow virtualization. In *BoF-Session on Data Citation* [14]. Research Data Alliance - Launch and First Plenary March 18-20, 2013, Gothenburg, Sweden.

[10] C.-I. T. G. on Data Citation Standards and Practices. Out of cite, out of mind: The current state of practice, policy, and technology for the citation of data. *Data Science Journal*, 12:CIDCR1–CIDCR75, 2013.

[11] H. Pampel, H. Pfeiffenberger, A. Schäfer, E. Smit, S. Pröll, D. Giaretta, and S. Lambert. D33.1A - Report on Peer Review of Research Data in Scholorly Communication. Technical report, APARSEN, 04 2012.

[12] N. Paskin. Digital Object Identifier (DOI) System. *Encyclopedia of library and information sciences*, 3:1586–1592, 2010.

[13] H. Pfeiffenberger, H. Pampel, A. Schäfer, V. Guidetti, C. Bruch, Y. Tzitzikas, S. Pröll, R. Mayer, S. Mele, P. Herterich, S. Dallmeier-Tiessen, M. Grootfeld, and R. van Horik. D26.1 - Report and Strategy on Annotation, Reputation and Data Quality. Technical report, APARSEN, 04 2012.

[14] S. Pröll and A. Rauber. BoF-Session on Data Citation. March 2013. Research Data Alliance - Launch and First Plenary March 18-20, 2013, Gothenburg, Sweden.

[15] S. Pröll and A. Rauber. Citable by Design - A Model for Making Data in Dynamic Environments Citable. In *2nd International Conference on Data Management Technologies and Applications (DATA2013)*, Reykjavik, Iceland, July 29-31 2013.

[16] S. Pröll and A. Rauber. Scalable Data Citation in Dynamic, Large Databases: Model and Reference Implementation. In *IEEE International Conference on Big Data 2013 (IEEE BigData 2013). Accepted for publication.*, 10 2013, Accepted for publication.

[17] S. Salza, M. Guercio, M. Grossi, S. Pröll, C. Stroumboulis, Y. Tzitzikas, M. Doerr, and G. Flouris. D24.1 Report on Authenticity and Plan for Interoperable Authenticity Evaluation System. Technical report, APARSEN, 04 2012.

[18] R. T. Snodgrass, J. Gray, and J. Melton. *Developing time-oriented database applications in SQL*, volume 42. Morgan Kaufmann Publishers San Francisco, 2000.

[19] R. M. S. S. R. V. J. B. Tomasz Miksa, Stefan Pröll and A. Rauber. Framework for verification of preserved and redeployed processes. In *Proceedings of the 10th International Conference on Digital Preservation (iPRES2013)*, 9 2013.