Repurposing Benchmark Corpora for Reconstructing Provenance

Sara Magliacane, Paul Groth

Department of Computer Science
VU University Amsterdam
s.magliacane@vu.nl, p.t.groth@vu.nl

Abstract. Provenance is a critical aspect in evaluating scientific output, yet, it is still often overlooked or not comprehensively produced by practitioners. This incomplete and partial nature of provenance has been recognized in the literature, which has led to the development of new methods for reconstructing missing provenance. Unfortunately, there is currently no agreed upon evaluation framework for testing these methods. Moreover, there is a paucity of datasets that these methods can be applied to. To begin to address this gap, we present a survey of existing benchmark corpora from other computer science communities that could be applied to evaluate provenance reconstruction techniques. The survey identifies, for each corpus, a mapping between the data available and common provenance concepts. In addition to their applicability to provenance reconstruction, we also argue that these corpora could be reused for other tasks pertaining to provenance.

1 Introduction

Data provenance, or the "history of data' [15], is an increasingly important aspect of data management in several settings, forming the foundation of trust, repeatability, attribution and metrics. In scholarly publishing, some aspects of provenance are already tracked manually, traditionally through the mechanisms of authorship and citations. Even though tracking other aspects of provenance would enable a more accurate description of the information flow in science, it is often neglected as a computationally and organizationally intensive task. Moreover, even if the tracking of provenance could be enforced for future publications, how can we include into this vision past and current "legacy" publications?

In most cases, in the absence of actual provenance, we may still prefer to have a plausible reconstruction of what may have happened. One possibility is to automatically generate some hypotheses based on the content of the data and possibly metadata, in a way that resembles forensic investigation, which tries to reconstruct a series of past events based on the available evidence. In this paper, we refer to this task as the problem of provenance reconstruction.

We consider as special cases of provenance reconstruction the following three use cases that are important in scholarly publishing:

– detecting plagiarism, text and multimedia content reuse;

– connecting publications with related data, both research data and other content (blog posts, presentations and videos);
– tracking the evolution of scientific knowledge and discourse through publications and informal communications between scientists;

A number of authors have presented techniques for reconstructing provenance [6,8,7,10,13,14]. However, each approach has been evaluated on different datasets and within different environments. For example, [6] focuses on extracting provenance from newspaper texts whereas [8] uses information from extensive logging within an operating system to create provenance traces. In the context of reconstructing provenance for the scholarly publishing process the related work is evaluated on manually annotated datasets: in particular, [13] describes an approach to reconstruct the provenance of a shared folder containing all the files related to a scientific paper, e.g. TEX files, images and other accessory files, while [14] focuses on the reconstruction of the relationships between a set of papers and clinical guidelines. In some cases, for privacy concerns, the data cannot be made available as is the case with [7]. Because of this heterogeneity, it is difficult to compare the various methods and approaches in a systematic fashion.

Among the openly available provenance datasets, e.g. the ones collected by the ProvBench initiative[1], most focus primarily on provenance generated from computational workflows and not on other environments. An exception is Wikipedia-PROV, a dataset containing the Wikipedia edits provenance graph[2]. Furthermore, often these datasets only provide provenance graphs and not the corresponding information that the provenance refers to. For provenance reconstruction, such information is vital as the reconstruction is based on the content. To evaluate reconstruction methods, one needs to have a gold standard provenance graph and the underlying data from which that gold standard can be built. To begin to address the paucity of datasets, we performed a survey of existing benchmark corpora from other computer science communities that could be applied to evaluate provenance reconstruction techniques. Thus, the paper makes the following contributions:

– a survey of existing benchmark corpora with respect to provenance reconstruction;
– an in-depth analysis of how two of these corpora can be mapped to the W3C PROV model of provenance [11].

More broadly this paper aims to contribute to the ongoing discussion around provenance benchmarks by identifying existing corpora that could be used for benchmark provenance reconstruction approaches. Additionally, it ask the question of how to integrate existing content with next generation publications.

The rest of this paper is organized as follows. We begin by describing our survey methodology and a review the corpora themselves. We then focus on two examples in-depth. Finally, we conclude with some observations about these

[1] https://sites.google.com/site/provbench/
[2] https://github.com/provbench/Wikipedia-PROV

datasets and their applicability to both the specific problem of provenance reconstruction and wider use-cases.

2 Methodology

To collect corpora, we did a focused search concentrating primarily on datasets that have already been used in various computer science evaluation initiatives (e.g TREC). Each corpus was analyzed for its usefulness with respect to provenance and in particular provenance reconstruction. This was done by identifying whether the data could represent information from one or more broad classes of provenance information. The three classes we used are identified below. For each class, we describe how provenance can be concretely expressed using concepts from the W3C PROV data model [16].

- Dependency - a dependency between two objects expressed as the relationship between two prov:Entity objects, e.g. prov:wasRevisionOf or prov:wasDerivedFrom;
- Sequence of operations - a process expressed as a sequence of prov:Activity that connect two prov:Entity objects, expressed through prov:used and prov:wasGeneratedBy relations;
- Authorship - attribution information expressed as the prov:Agent that created the Entity using the prov:wasAttributedTo relation.

These classes reflect the three use-case perspectives on provenance identified by the W3C Provenance Primer [9]: object-oriented, process-oriented and agent-oriented. Thus, these classifications should help guide researchers to useful datasets depending on the perspective their technique is intended for. A key heuristic that we used when deciding whether to incorporate a dataset in the survey was whether it could be used to express not just similarities but dependencies.

3 Survey of existing corpora

There are several available corpora in the Natural Language Processing and Information Retrieval communities. In line with their tasks, most of them provide information about the relevance of entities for a given query or similarity between entities, fewer provide information dependency or influence relationships between entities necessary to act as provenance. Among the existing benchmark corpora that contain provenance-like information most are text-based with only a few containing image and video data.

3.1 Text corpora

Plagiarism detection and text reuse [4] are two related and established fields that can be seen as a special case of reconstructing provenance, especially dependencies between entities. These also play an important role with respect to

scientific literature. Textual entailment can also be seen as a special case of sentence-level provenance. Finally, citation networks provide, what can be seen as, a provenance graph of publications. The following datasets come from these areas.

1. **Corpus Name:** METER corpus [5]
   Availability: Available after registration.[3]
   Background: A journalistic text reuse corpus, consisting of a set of news stories from the major UK news agency and the related news items from nine British newspapers.
   Content: 445 cases of text reuse in 1,716 text documents, annotated by a domain expert in terms of how much the newspaper stories were derived from the agency story and whether there had been some word or phrase reuse. We note that the data was annotated by one, albeit expert, annotator, which could impact upon the accuracy of the information.
   Relationship to Provenance: This corpus can be seen as describing both the dependency and the sequence of operations, reduced to the two basic activities of word reuse and phrase reuse, across the news stories. On the other side, the considered relationships are always from an agency story to a newspaper story, not between agency stories or between newspaper stories.

2. **Corpus Name:** PAN-PC-12 detailed comparison training corpus (an improved version of the PAN-PC-10 [19])
   Availability: Directly available.[4]
   Background: Used in the Plagiarism detection (PAN) 2012 competition[5] in the detailed comparison task.
   Content: The corpus contains 4,210 source documents, derived from the books of Project Gutenberg, and 1,804 "suspicious" documents, where "suspicious" means that they may or may not contained one or more plagiarized passages. In total there are 5,000 plagiarism cases. Each plagiarized passage is annotated with the source passage in the source document. The plagiarism cases were either simulated by crowd-sourcing the rewriting and paraphrasing, or generated artificially through three obfuscation strategies: Random Text Operations (shuffling, removing, inserting or replacing words at random), Semantic Word Variation (replacing word by their synonyms, hyponyms, etc.) and POS-preserving Word Shuffling (shuffling words at random while retaining the original part-of-speech sequence).
   Moreover, there are cases of cross-language plagiarism, which in the past editions of the competition [19] were constructed by applying Google Translate. In PAN-PC-12 they are generated based on the multilingual Europarl corpus [12] by inserting the English version of an originally German or Spanish passage into a Gutenberg book.
   Relationship to Provenance: The released corpus contains information

[3] http://nlp.shef.ac.uk/meter/
[4] http://www.webis.de/research/corpora/corpus-pan-pc-12/pan12/
[5] http://pan.webis.de

on the dependency between entities, in this case paragraphs. This could be improved by tracking the performed operations of the process of automatically generating the corpus. If this was possible, the corpus could also be used as a record of a sequences of operations.

3. **Corpus Name:** Wikipedia co-derivative corpus [2]
   Availability: Available after registration.[6]
   Background: A corpus based on Wikipedia edit history.
   Content: 20,000 documents in four languages (German, English, Hindi and Spanish). For each language, the top 500 most popular Wikipedia articles are retrieved, each with ten revisions.
   Relationship to Provenance: The ten revisions of each article are connected by an edit activity, therefore the corpus contains dependencies between entities. On the other side, the activity is not characterized in more detail, but is just marked as an "edit" operation.

4. **Corpus Name:** PAN-WVC-11 (an improved version on PAN-WVC-10 [18])
   Availability: Directly available.[7]
   Background: Used in the PAN 2011 competition in the Wikipedia vandalism task.
   Content: 29,949 edits on 24,351 Wikipedia articles in three languages (9,985 English edits, 9,990 German edits, and 9,974 Spanish edits), among which 2,813 edits are vandalism edits. The annotated corpus has been crowdsourced using Amazon's Mechanical Turk.
   Relationship to Provenance: This corpus can be thought of as a basic sequence of operations with only one activity, which can be either a legitimate edit or a vandalism edit.

5. **Corpus Name:** PAN-AI-11 training datasets [1]
   Availability: Directly available.[8]
   Background: Used in the Authorship identification task of the PAN 2011 competition, based on a subset of the Enron email dataset.
   Content: More than 12,000 emails written by 118 Enron managers, divided in two subsets:"Large", containing 9337 document by 72 authors, and "Small", containing 3001 documents from 26 authors. The emails were attributed based on the "From: " headers, and multiple emails were reconnected to the same author.
   Relationship to Provenance: The author can be seen as the agent that performs an activity on the document.

[6] http://users.dsic.upv.es/grupos/nle/resources/abc/download-coderiv.html
[7] http://www.uni-weimar.de/cms/medien/webis/research/corpora/corpus-pan-wvc-11.html
[8] http://www.uni-weimar.de/medien/webis/research/events/pan-11/pan11-web/author-identification.html

6. **Corpus Name:** MLaF at FIRE 2011
   Availability: Available after signing the FIRE agreement.[9]
   Background: Used in the Mailing Lists and Forums Track at the FIRE 2011 competition, where the task was the classification of messages from mailing lists and forum discussions in a set of seven types of message.
   Content: 212 132 documents from ubuntu-users (from September 2004 to June 2009) and tugindia (May 2001 - June 2009) mailing lists and on several technical and tech support forums. The corpus maintained the natural causal ordering of the messages in the forums and the threads in the mailing list using the "In-reply-to" fields. Each of the messages is classified as belonging to one or more of seven predetermined categories: e.g. ASK_QUESTION, ASK_CLARIFICATION and SUGGEST_SOLUTION.
   Relationship to Provenance: The dependency relationship is inherent in the structure of the threads, moreover, these categories reflect the activity that generated the messages.

7. **Corpus Name:** RTE-7 [3]
   Availability: Available after signing the Past TAC data agreement[10]. Some older versions (e.g. RTE-3) are available directly.[11]
   Background: Used in the RTE (Recognizing Textual Entailment)[12] challenge at TAC 2011. The main task consisted in determining whether one text fragment is entailed, i.e. can be inferred, from another.
   Content: The corpus contains: a development set of text fragments with 10 topics, 284 hypotheses and 21,420 candidate entailments, of which 1 136 are judged as correct, and a test set with 10 topics, 269 hypotheses and 22,426 candidate entailments, of which 1,308 are judged as correct. The text fragments are based on the TAC 2008 and 2009 Update Summarization Task and the entailment was annotated by three annotators.
   Relationship to Provenance: Textual entailment can be seen as a form of dependency among text fragments. Unfortunately, the activity connecting these fragments cannot be further characterized beyond the entailment.

8. **Corpus Name:** arXiv HEP-PH citation graph from KDD 2003
   Availability: Directly available.[13]
   Background: The articles and citation graph of the high energy physics phenomenology articles uploaded to arXiv between January 1993 and March 2003 from the Citation Prediction task of the 2003 KDD Cup
   Content: 34,546 papers that contain 421,578 references, some of which refer to publications outside of the dataset. The dataset includes the LaTeX source of the main .tex file and several arXiv metadata, like the submission and revision dates, the authors and abstract. Since some of the articles were

[9] https://sites.google.com/site/mlaffire/the-data
[10] http://www.nist.gov/tac/data/past/2011/RTE-7_Main_Task.html
[11] http://pascallin.ecs.soton.ac.uk/Challenges/RTE3/Datasets/
[12] http://www.nist.gov/tac/2011/RTE/
[13] http://www.cs.cornell.edu/projects/kddcup/

older than the submission, it also contains the original publication date.
Relationship to Provenance: The citation networks represent the dependency between publications. The networks not only consider text reuse and paraphrasing, but also textual entailment and summarization. On the other side, if there are any plagiarism cases, they are unlikely to cite the original article so the data may be incomplete with respect to provenance.

3.2 Image corpora

To the best of our knowledge, there is no competition for image reuse or image copy detection, although there is extensive literature on the subject (e.g. see [21] for a comparison of possible approaches). Therefore it was difficult to find publicly available corpora with ground truth annotations that could be repurposed for provenance. The most related competition corpus that we found is used for event detection.

1. **Corpus Name:** Social Event Detection 2012 (SED 2012) dataset [17]
   Availability: Directly available.[14]
   Background: Used at the MediaEval 2012[15] competition in the Social Event Detection task. The task consists in detecting social events and finding clusters of images related to each event. There are three challenges, each related to a specific kind of event, for example the first challenge is "Find technical events that took place in Germany in the test collection".
   Content: 167,332 images captured between the beginning of 2009 and end of 2011 by 4,422 unique Flickr users. For each image there are some metadata available (e.g. time-stamps, tags, 20% of the images have also geotags). The images ewre collected using queries for specific events on the Flickr API.
   Relationship to Provenance: The images of each cluster are all related to the same event, therefore there is a dependency between them.

3.3 Video corpora

Among video retrieval competitions, the most relevant task for our work is the copy detection task. This task has been present at the TRECVid since 2008, but in this paper we describe only the 2011 dataset.

1. **Corpus Name:** CCD at TRECVID 2010[16].
   Availability: Available after signing the TRECVid agreement. Another possibility is to recreate the dataset using the provided tools.
   Background: Used at the TREC Video Retrieval Evaluation (TRECVid)[20] 2010 competition in the content-based copy detection task (CCD). A copy is a segment of video derived from another video using some transformations.

[14] http://mklab.iti.gr/project/sed2012/
[15] http://www.multimediaeval.org/
[16] http://www-nlpir.nist.gov/projects/tv2010/#ccd

Content: The corpus is based on two reference datasets of videos: IACC.1.A, which contains about 8000 Internet Archive videos (MPEG-4 H.264, 50GB, 200 hours) with duration between 10 seconds and 3.5 minutes, and IACC.1.-tv10.training, which contains about 3200 Internet Archive videos (50GB, 200 hours) of around 4 minutes. For most videos the metadata are also available. The queries are constructed from the reference data using specific tools that apply one or more transformations from a known set. This set includes inserting patterns, compression, picture in picture (a video inserted in the front of another video) and post production transformations (e.g. crop, shift, flip).
Relationship to Provenance: There is a dependency between each couple of original and copied video segments that is realized through a sequence of activities, chosen from a known set of transformations.

3.4 Summary of the survey

In Table 1, we present a summary of the surveyed corpora, where each corpus is classified based on the type of data it represents, the operations that are tracked and the information about the authors. In this classification, we did not consider dependencies, because all of the surveyed corpora cover this aspect. An empty cell in the table represents the fact that there is no information regarding to that aspect of provenance. In the case of operations, it means that there are no explicit operations tracked. From Table 1, we can see that for sequences of operations the most promising datasets are PAN-PC-12 [19], MLaF at FIRE 2011 and CCD at TRECVID 2010[20], due to the different categories of operations they capture.

4 Two example corpora

We now look at two corpora in more detail, but the considerations and methods we use could be extended to the other corpora. As a representative of the text corpora, we chose the corpus from the Plagiarism Detection competition (PAN) [19], which provides the most natural and interesting setting for provenance reconstruction. In the view of the increasing multimedia nature of scientific publications, we consider also other forms of content reuse, in particular video content reuse. As a representative of the video corpora, we chose the corpus of TRECVid [20], a well-established competition for video information retrieval and a very good example of sequences of operations reconstruction. For each of these corpora we propose a conversion to the PROV standard, which allows it to be used in a variety of existing applications. Moreover, this conversion enables the connection between these corpora and other provenance datasets.

4.1 Text corpus: PAN-PC-12

We first consider the PAN-PC-12 corpus, which is an updated version of PAN-PC-10 [19]. From this corpus, we consider the detailed comparison corpus, which associates each plagiarized paragraph with the source paragraph. The dataset

Corpus	Type	Operations	Authorship
METER [5]	Newspaper articles (text)	Word reuse, Phrase reuse	-
PAN-PC-12 [19]	Plagiarized books (text)	5 types of plagiarism	-
Wikipedia co-derivative [2]	Wikipedia edits in 4 languages (text)	Edit	-
PAN-WVC-11 [18]	Wikipedia vandalized edits in 3 languages (text)	Edit, vandalization	-
PAN-AI-11 [1]	Emails from 118 authors (text)	-	Email authors
MLaF at FIRE 2011	Mailing lists and forum discussions(text)	7 categories of answer	-
RTE-7 [3]	Text fragments and entailments (text)	-	-
arXiv HEP-PH at KDD 2003	Scientific publications (text)	Cite	Authors
SED-2012	Images about social events (images, tags)	-	-
CCD at TRECVID 2010[20]	Video content reuse examples (video)	10 transformations	-

Table 1: Summary of the survey

contains all the documents in text format and one XML file per document that contains some metadata like author, title and language. Moreover, in the case of suspicious documents, it describes for each plagiarized paragraph the offset and length of the source paragraph, and the source document. The dataset distinguishes several types of plagiarism:

- artificial plagiarism with high/low/no obfuscation;
- translated plagiarism;
- simulated plagiarism with paraphrase (crowd-sourced).

We propose to convert the dataset to a PROV template similar to the one presented in Figure 1. In particular, we model a suspicious document as a collection of several paragraphs, some of which are original and some of which are a result of plagiarism. The plagiarized paragraphs are derived from the original paragraphs through a Plagiarism activity of any of the above-mentioned five types. In addition, each of the original paragraphs is contained in an original document.

4.2 Multimedia corpus: TRECVid 2010

The corpus of the content-based copy detection task of the TREC Video Retrieval Evaluation [20] is a good example multimedia corpus for provenance

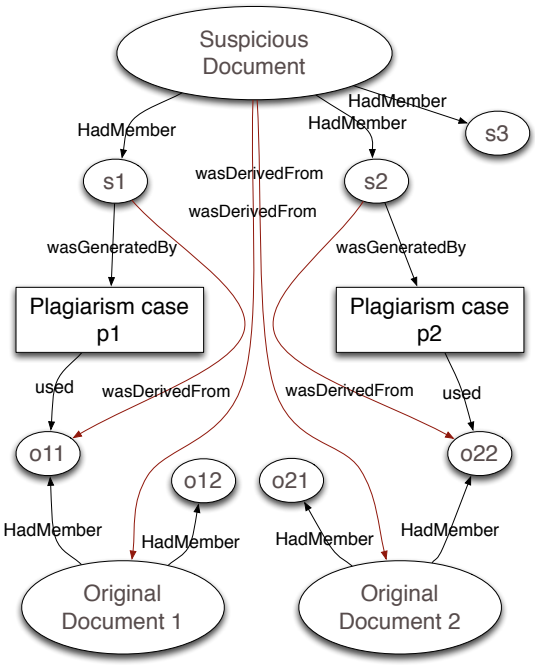

Fig. 1: The PROV template for PAN-PC-12

reconstruction. In the TRECVid terminology, the queries are the plagiarized copies, which are constructed from the two datasets by applying one or more transformation from a known set of ten transformations. Each of the transformations has one or two input videos and several numerical parameters. The transformations can be:

- *basic*, e.g. cam-coding, insertion of a pattern, picture in picture (a video is inserted in the front of another video), blur, crop, shift;
- *composed*, i.e. sequences of three or five basic transformations.

We propose to convert the dataset to a PROV template similar to the one presented in Figure 2. In this case, the generated video is created as an output of the Transformation activity, that can be characterized as one of the types of transformations with certain parameters. The inputs are one or two videos from the reference collection.

5 Analysis & Conclusion

Overall, these corpora provide a good test sets for provenance systems focused on the agent or entity oriented perspectives. However, none of these corpora provide information that can be construed as provenance between different types

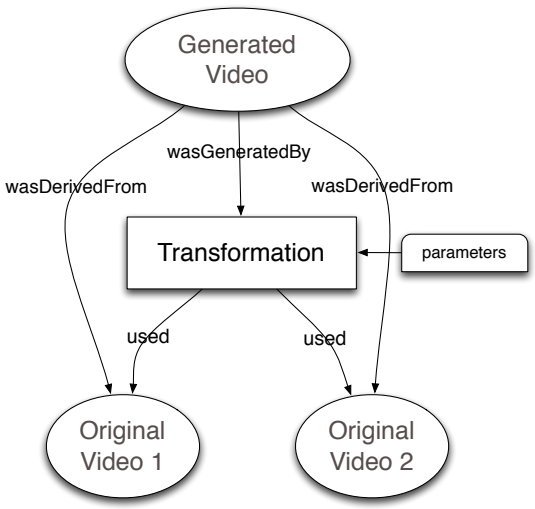

Fig. 2: The PROV template for TRECVid

of entries (e.g. text+image) or as representing long complex and open chains of activities. Additionally, the datasets clearly do not cover all the types of information regularly defined in provenance models, for example, rich semantics about the types of agents or activities within a provenance trace. However, we believe they provide a usable testing environment, in particular, for the reconstruction task. These datasets may also prove useful in systems that want to test the scalability of their provenance infrastructure because of the large amounts of data involved in some of the corpora, for example, TRECVID.

To conclude, in this paper, we provided an overview of existing benchmark corpora that could be used for testing provenance systems and in particular provenance reconstruction methods. We believe that these existing datasets provide a good first step for testing such systems. We hope to provide extracted provenance graphs from a select set of the surveyed datasets. However, going forward, there is a clear need for both manually curated and synthetic provenance-specific benchmarks. The ability to reconstruct provenance will be a key part of integrating existing content into the next generation of scientific publications.

Acknowledgements This publication was supported by the Data2Semantics project in the Dutch national program COMMIT.

References

1. Argamon, S., Juola, P.: Overview of the international authorship identification competition at pan-2011. In: CLEF 2011 (2011)

2. Barrón-Cedeño, A., Eiselt, A., Rosso, P.: Monolingual Text Similarity Measures: A Comparison of Models over Wikipedia Articles Revisions. pp. 29–38
3. Bentivogli, L., Clark, P., Dagan, I.: The seventh pascal recognizing textual entailment challenge. Text Analysis Conference (TAC) 2011 Notebook Proceedings (2011)
4. Broder, A.Z.: On the resemblance and containment of documents. In: In Compression and Complexity of Sequences (SEQUENCES'97 (1997)
5. Clough, P., Gaizauskas, R., Piao, S.: Building and annotating a corpus for the study of journalistic text reuse. LREC 202 (2002)
6. de Nies, T., Coppens, S., van Deursen, D.: Automatic Discovery of High-Level Provenance using Semantic Similarity. IPAW 2012 (2012)
7. Deolalikar, V., Laffitte, H.: Provenance as data mining: combining file system metadata with content analysis. In: First workshop on on Theory and practice of provenance. p. 10. USENIX Association (2009)
8. Frew, J., Metzger, D., Slaughter, P.: Automatic capture and reconstruction of computational provenance. Concurrency and Computation: Practice and Experience 20(5), 485–496 (2008)
9. Gil, Y., Miles, S., Belhajjame, K., Deus, H., Garijo, D., Klyne, G., Missier, P., Soiland-Reyes, S., Zednik, S.: A primer for the prov provenance model (2012), http://www.w3.org/TR/prov-primer/, world Wide Web (W3C)
10. Groth, P., Gil, Y., Magliacane, S.: Automatic Metadata Annotation through Reconstructing Provenance. In: Semantic Web in Provenance Managment workshop (2012)
11. Groth, P., Moreau, L.: http://www.w3.org/TR/prov-overview/
12. Koehn, P.: Europarl: A Parallel Corpus for Statistical Machine Translation. In: Machine Translation Summit X. pp. 79–86. Phuket, Thailand (2005)
13. Magliacane, S.: Reconstructing provenance. The Semantic Web–ISWC 2012 (2012)
14. Magliacane, S., Groth, P.: Towards Reconstructing the Provenance of Clinical Guidelines. In: Proceedings of the 5th International Workshop on Semantic Web Applications and Tools for Life Sciences (SWAT4LS). CEUR Workshop Proceedings, vol. 952. Paris, France (2012)
15. Moreau, L.: The Foundations for Provenance on the Web. Foundations and Trends® in Web Science 2(2-3), 99–241 (2010)
16. Moreau, L., Missier, P.: PROV-DM: The PROV Data Model, http://www.w3.org/TR/prov-dm/
17. Papadopoulos, S., Schinas, E., Mezaris, V., , Troncy, R., Kompatsiaris, I.: Social Event Detection at MediaEval 2012 : Challenges , Dataset and Evaluation. MediaEval 2012 Workshop (2012)
18. Potthast, M.: Crowdsourcing a Wikipedia vandalism corpus. Proceedings of the 33rd international ACM SIGIR 2010 pp. 7–8 (2010)
19. Potthast, M., Stein, B.: An evaluation framework for plagiarism detection. Proceedings of the 23rd International Conference on Computational Linguistics: Posters (2010)
20. Smeaton, A.F., Over, P., Kraaij, W.: Evaluation campaigns and trecvid. In: MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval. pp. 321–330. ACM Press, New York, NY, USA (2006)
21. Thomee, B., Huiskes, M.J., Bakker, E.M., Lew, M.S.: Large scale image copy detection evaluation. In: Lew, M.S., Bimbo, A.D., Bakker, E.M. (eds.) Multimedia Information Retrieval. pp. 59–66. ACM (2008)