Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 18 (2013) 1172 – 1178

Procedia Computer Science

2013 International Conference on Computational Science

Automatic hypothesis checking using eScience research infrastructures, ontologies, and linked data: a case study in climate change research

Jaakko Lappalainen[*†], Miguel-Ángel Sicilia[‡], Bernabé Hernández[§]

Computer Science department, University of Alcalá, Polytechnic Building, Ctra. Barcelona Km. 33.6. 28871 Alcalá de Henares, Spain

Abstract

Current research on physical or natural phenomena produces large datasets used by scientists to prove or disprove hypotheses. The increasing detail of our science triggers the trend of Big Data, and introduces new challenges on managing and processing that information. Allowing future reference to the data that generated previous research is a fundamental need and key feature of the scientific method, as experiments have to be reproduce efficiently by the research community. In this paper, we present a novel approach to manage such a large dataset in the climate research domain. With the help of ontologies and a Linked Data approach to describe the dataset, we allow computers to automatically prove hypothesis as new data of the same phenomena is integrated, realizing the vision of automated "executable research" that uses computational resources for continuously running exploratory hypotheses.

Keywords: eScience; ontologies; linked data; grid; big data;

1. Introduction

Physical and natural sciences are pushing the new trend of Big Data, a new data management paradigm that requires extremely large data collections to be processed over high performance computer systems, and consequently, generate large data as a result that cannot be handled by conventional databases.
In almost every natural science field, a computational specialization has appeared years ago. From particle

* Corresponding author. Tel.: +34-91-885 6518 ; fax: +34-91-885-6518 .
E-mail address:jkk.lapp@uah.es.

‡ msicilia@uah.es
§ bernabe_hernandez_hernandez@hotmail.com

doi:10.1016/j.procs.2013.05.283

science to the modern computational biology, science is becoming more and more dependent and jeopardized by technical limitations of current computing systems. But the data deluge is also providing new opportunities for scientific discovery, notably in the direction of the paradigm of data-intensive science [22, 25, 26, 27].
Observation of nature consists on harvesting samples of physical or natural events, and more detailed observations require bigger storage systems. Examples of this kind of data are astronomical images, nuclear physics experiment sensor data, and other raw data to be analyzed.
Simulation is another family of scientific computation approaches in which researchers aim to reproduce some natural phenomenon following a mathematical model that tries to predict its behavior. The results of the simulations are the relevant numerical parameter values the scientist is interested in.
Either by observing the nature or simulating it, post-processing is necessary to analyze the data gathered. Most common post-processing are based on statistical analysis and pattern recognition, in which tailored mathematical models may fit. A remarkable post processing activity often performed in computational science is visualization that helps scientist to understand physical laws underlying the studied event.
In the previous mentioned activities, as the detail of our science increases, so does the size of data, which triggers new challenges on computer science fields of data management and processing. This is also bringing new opportunities for exploring and re-evaluating hypotheses and theories. Typically researchers focus on a particular time frame and scope for testing their hypotheses. The subsequent report of the result, often in the form of a scholarly publication, then is a "snapshot" limited in scope and time coverage. However, in data-driven science, data relevant to some hypotheses gets continuously aggregated as time passes, and eventually, with the help of common semantics, it can be combined or related to other datasets. In this new context, it makes sense to represent the hypothesis as programs that are executed repeatedly, as new relevant amounts of data gets aggregated. This ability to automatically test hypotheses can also be applied to "exploratory" data science, in which a set of hypothesis is represented as a program that gets executed on different combination of variables and is run in a process of discovery that may eventually trigger some result that would be worth examining by the researchers. This idea has antecedents in knowledge discovery techniques as association rule mining [23], but it is different in scope, broader and theory informed and requires shared dataset semantics in advance.
The rest of this paper is structured as follows. In Section 2 we expose our motivation to develop this work. Section 3 gathers related work in fields of computational science, automatic scientific publication processing and big data. Section 4 presents the case study we have conducted, and section 5 shows empirical results and relevant quantitative information of our experiments. The paper ends with section 6, in which we summarize our work, provide conclusions and future work.

2. Motivation

The scientific method demands from the science practitioner to describe the process and protocols used in his work and provide the materials and results to the research community to become acknowledged. This enables the community to peer-review not only the final paper to be accepted in a specific conference or journal, but more globally, allows the entire community interested in that field to reproduce and check the correctness, accuracy and validity of the research. In science, this is a simplified example of how new data-intensive science is adopted.

But in order to share this science, and specifically Big Data, and to make it reusable to scientists a large amount of resources is needed to access and manipulate that very data. A first naïve approach to this is to establish powerful communication and transportation devices to transmit data as fast as possible, but this does not scale when scientists always pursue to exploit resources over interactive or online paradigms.

This is the reason why it is necessary to describe properly research data and allow machines to automatically detect new data to be integrated and checked to prove or disprove previous research hypothesis and theories. Then, hypotheses' checking becomes a process rather than a single step in a research plan, and hypotheses are put to the test continuously as new data arrives. Obviously, this requires two main elements: (a) semantic

integration of datasets to enable their integration in order to combine them in the same hypothesis checking process and (b) computational support for the automated execution. This enables a process of "hypothesis or conjecturing as a service" that provides a new dimension to scientific inquiry. The computational support required can be implemented in the form of Data Grids. Recent studies in this field appeared, such as [28, 29, 30]. These Grids are often national or international joint research projects, to serve a large amount of scientists from different countries. The authors of this project are linked to the agINFRA project [34], and use this data infrastructure to support the present work. Regarding semantics, the synergy of standards for publishing dataset descriptions as DCAT [24] with approaches to semantics using Linked Open Data provide a practical and increasingly adopted framework for the concept presented in this paper.

3. Related Work

There have been works on integrating science data, such as in [1], authors show how semantic extensions facilitate complex data integration, when the data covers different but related science fields. An approach to develop a standard ontology for biomedical sciences is described in [4], to give an answer to the proliferation of different ontologies in this field that established obstacles to data integration. Authors in [5], present a novel bioinformatics data warehouse software kit that integrates biological information from multiple public life science data sources into a local database management system. In [6], authors propose a formal foundation for merging XML-like data and discuss indexing support for data merging.

For managing large scientific data on grid and cloud systems, [10] examines some of approaches to manage this data on its entire lifecycle. Works like [7] try to cluster datasets hierarchically, by designing an algorithm that uses a multidimensional grid data structure to organize the value space surrounding the pattern values, rather than to organize the patterns themselves. In [8], authors describe a framework designed to provide support for splitting and processing of datasets in a distributed and heterogeneous environment. As for [9], authors describe the architecture and implementation of a system built to support batch processing of large scientific datasets, over the Internet. This system implements a federation of autonomous workstation pools, which may be widely distributed. Individual batch jobs are executed using idle workstations in these pools. In [14], authors use the Hadoop cloud-computing framework to develop a user application that allows processing of scientific data on clouds.

There is a survey [13] on scientific data repositories, which store and describe scientific data using ontologies and linked data; other descriptions and ontologies have been done in space science by authors in [2], which developed a set of solar-terrestrial ontologies as formal encodings of the knowledge in the Ontology Web Language–Description Logic [3] (OWL–DL) format. In [11], authors describe a linked data search engine for scientific data is presented. It has been used in the Scientific Database project of Chinese Academy of Sciences. The paper describes the workflow, consisting of four processes including publishing, fetching scientific data and metadata, searching scientific data and discovering links among them. For [12], the authors present an ontology-centric data management system architecture for scientific experimental data that is extensible and domain independent. In this architecture, the behaviors of domain concepts and objects are specified entirely by ontological entities, around which all data management tasks are carried out. The open and semantic nature of ontology languages also makes the architecture amenable to greater data reuse and interoperability.

Automatic scientific data processing and scientific publication treatment has been used by [15, 16, 17], to develop systems for automatic extraction of metadata from scientific papers in PDF format for the information system for monitoring the scientific research activity. In [18], a tool is describe to assist on paper writing, automatically analyses the publication, evaluates consistency parameters and interactively delivers feedback to the author. It analyses the proper use of acronyms and their definitions, and the use of specialized terminology. It provides Gene Ontology (GO) and Medline Subject Headings (MeSH) categorization of text passages, the

retrieval of relevant publications from public scientific literature repositories, and the identification of missing or unused references.

4. Case study

To illustrate the method and techniques explored in this paper, we show a case study on climate change research. The starting point of this work is a paper published in the early 2000's which focuses on the temperature series trends in Australia during the 20th century.

4.1 The hypothesis to be tested

In 2002, Lenten and Moosa published a work [19] using econometric time series tools to study the trends on Australian surface air temperatures in the 20th century. The conclusion the authors reached, is that the time series was I(1). This is the perfect example for the authors of this paper to put in practice the approach of "live science" introduced in previous sections.
The work of Lenten and Moosa is frozen in time, specifically from January 1901 to December 1998, and in space, using only data from 6 weather stations among Australia. But with advance computation resources and techniques, their hypothesis can be extended and automatically tested not just since 2002 to current date, but also in the future, and to other Australian territory, as we will show. The authors use temperature data from 6 weather stations in Australia, and conclude that there is an upward trend in many of those stations. Authors also express a concern of errors on the data considered. For missing data, authors in [19] use Kalman filters, but we can address this problem querying other datasets, as they get integrated.
This and other issues show the necessity of using more than one dataset to test the hypothesis, which is another reason we chose to gather data from two datasets.

4.2 Datasets to test by

The National Oceanographic and Atmospheric Administration (NOAA) is a U.S. federal agency that focuses its interest on marine and weather information within the country and outside. It gathers weather observation data from many stations all over the globe, including the ones we are interested in: the dataset on surface air temperature in the Australian territory. This dataset is daily updated in an FTP server[**] and it is publicly available, but not in LD: the information is in plain text compressed by months, and by each station. It is necessary a process of 'triplification' of this information, to enable processes and computations easily, allowing semantic web applications to access this data.
The other source of temperature data we consider is the one published by the Australian Bureau of Meteorology. As opposed to the data from NOAA, the ACORN-SAT dataset[††] is published in Linked Data using an ontology developed by CSIRO to this purpose.
Using two datasets representing the same phenomena, we can expand horizontally the hypothesis testing power, as an attempt to overcome the limitations of one dataset it terms of errors, missing values, and so on. But handling two datasets of this size is impossible without parallel computation systems, such as Grid and supercomputing environments.

4.3 Technical resources

[**] ftp.ncdc.noaa.gov
[††] http://lab.environment.data.gov.au

In this section we describe the technical resources used to complete this work.

4.3.1 Web crawler & Data fetcher

One important tool we used in the context of this work is the web crawler. A web crawler is a program that simulates the web dialogue with a web server and the client browser, in order to retrieve data.
The data from NOAA is not in LD, but it is very easy to write a JAVA program to fetch past data and ingest them to a relational database, and another program to daily fetch new data as it is provided. This makes possible to keep an updated dataset.

4.3.2 Grid infrastructure

The computational resource by which we retrieve and store science information is the one provided in the context of EU-funded agINFRA project [34]. With this data infrastructure, we can make unstructured data available in machine-readable format, and efficiently manage and process large datasets using a Linked Data paradigm.

4.3.3 Scientific routines to support hypothesis testing

The conclusion we want to test periodically is the one in Lenten and Moore's paper, which is that the temperature series is I(1). There are two important tests for this in the fields of statistics. The first one is the Phillips-Perron test [20], and the second is Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test [21], both are available to execute in R.

4.4 Getting and triplifying the data

Our data fetcher as an agINFRA application, can be executed in parallel. The computation and network power of infrastructure takes place in two ways. In the first one, the infrastructure executes the data fetcher daily in parallel. Each job fetches data from one weather station, and writes the new data to the relational database. The second task the infrastructure takes over is the triplification of that data, so it can be exposed as LD. The weather information in NOAA becomes a dataset in agINFRA, which can be processed by other users easily, as it is richly described with standard ontologies. In this case, we use SWEET [31], a well-known ontology developed in NASA JPL.

The web crawler we use is also an agINFRA application. It runs using 10 threads to crawl the URL of ACORN-SAT dataset. The crawler we implement is based on LDSpider [18], a framework for developing web crawlers for Linked Data. The output is produced in triple format, and easily ingested to a database.
To make the data openly accessible we publish the data from the NOAA database using a D2R [32] instance. This allows users to query data and make computations easily using a SPARQL client, as we show in the next section.

4.5 Processing the data with R

To prove the hypothesis in the paper [19], we need several packages available in R. First, we need a library to make HTTP requests to our SPARQL endpoints and be able to gather data. This is achieved using the package SPARQL [33]. The other important package we need is *tseries*, a mathematical library to execute the KPSS and PP tests, that proves the hypothesis of the unit root in a time series.

We have decided to schedule the execution of these scripts weekly, and provide the raw output of both tests into LD using a tailored vocabulary in the context of agINFRA data infrastructure.

It is now easy to write a semantic web application to query the result of the mathematical tests on the datasets.

5. Conclusions and future work

Currently, the scientific work over "live datasets" is static on time. The results presented in this work contrasts in their static nature against the evolving nature of the data that produces them. We allow hypotheses in scientific publications to be continuously checked as the amount of data become increases on time.

We have also shown very easily the integration a LD dataset in the case of CSIRO's temperature dataset, and the *triplification of* a non-structured dataset in the case of NOAA's database, using a parallel computing environment as agINFRA data infrastructure. In this scenario, agINFRA provides powerful computation and rich semantics to support traditional scientific workflows for natural sciences.

During the development of this work, new challenges have been triggered. In the integration of datasets, an interesting geographical mapping of weather stations needs be done to provide more complete and plural information in the climate change scenario. Also, to push further the automatic exploratory science on numerical datasets, a good starting point is the development of a core ontology to describe periodic hypothesis testing.

Experimenting with these techniques, it has been clear to the authors the limitations on automatic hypothesis checking. These limitations are similar to the ones encountered on classic NLP problems, as the scientific language is a subset of the natural language. The expressive power of current semantic technologies allows us to describe simple hypothesis such as trending analysis, correlation checking and some mathematical formula evaluation.

References

1. Ludäscher, B., Lin, K., Bowers, S., Jaeger-Frank, E., Brodaric, B., & Baru, C. (2006). Managing scientific data: From data integration to scientific workflows. *Geoinformatics: Data to knowledge*, *397*, 109.
2. Fox, P., McGuinness, D. L., Cinquini, L., West, P., Garcia, J., Benedict, J. L., & Middleton, D. (2009). Ontology-supported scientific data frameworks: The virtual solar-terrestrial observatory experience. *Computers & Geosciences*, *35*(4), 724-738.
3. McGuinness, D. L., & Van Harmelen, F. (2004). OWL web ontology language overview. *W3C recommendation*, *10*(2004-03), 10.
4. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., ... & Lewis, S. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, *25*(11), 1251-1255.
5. Töpel, T., Kormeier, B., Klassen, A., & Hofestädt, R. (2008). BioDWH: a data warehouse kit for life science data integration. *Journal of integrative bioinformatics*, *5*(2), 93.
6. Pankowski, T., & Hunt, E. (2005). Data merging in life science data integration systems. *Intelligent Information Processing and Web Mining*, 279-288.
7. Schikuta, E. (1996, August). Grid-clustering: An efficient hierarchical clustering method for very large data sets. In *Pattern Recognition, 1996., Proceedings of the 13th International Conference on* (Vol. 2, pp. 101-105). IEEE.
8. Beynon, M. D., Kurc, T., Catalyurek, U., Chang, C., Sussman, A., & Saltz, J. (2001). Distributed processing of very large datasets with DataCutter. *Parallel Computing*, *27*(11), 1457-1478.
9. Chen, C., Salem, K., & Livny, M. (1996, May). The DEC: processing scientific data over the Internet. In *Distributed Computing Systems, 1996., Proceedings of the 16th International Conference on* (pp. 673-679). IEEE.
10. Deelman, E., & Chervenak, A. (2008, May). Data management challenges of data-intensive scientific workflows. In *Cluster Computing and the Grid, 2008. CCGRID'08. 8th IEEE International Symposium on* (pp. 687-692). IEEE.
11. Shen, Z., Hou, Y., Li, C., & Li, J. (2012, May). Voovle: A linked data search engine for scientific data. In *Fuzzy Systems and Knowledge Discovery (FSKD), 2012 9th International Conference on* (pp. 1171-1175). IEEE.

12. Li, Y. F., Kennedy, G., Davies, F., & Hunter, J. (2010). Towards a semantic & domain-agnostic scientific data management system. In *The 9th International Semantic Web Conference (ISWC2010)* (pp. 13-24). Semantic Web Science Association.
13. Marcial, L. H., & Hemminger, B. M. (2010). Scientific data repositories on the Web: An initial survey. *Journal of the American Society for Information Science and Technology*, *61*(10), 2029-2048.
14. Zhang, C., De Sterck, H., Aboulnaga, A., Djambazian, H., & Sladek, R. (2010). Case study of scientific data processing on a cloud using hadoop. In *High performance computing systems and applications* (pp. 400-415). Springer Berlin/Heidelberg.
15. Kovacevic, A., Ivanovic, D., Milosavljevic, B., Konjovic, Z., & Surla, D. (2011). Automatic extraction of metadata from scientific publications for CRIS systems. *Program: electronic library and information systems*, *45*(4), 376-396.
16. Lopez, P. (2009). GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications. *Research and Advanced Technology for Digital Libraries*, 473-474.
17. Kim, S. N., Medelyan, O., Kan, M. Y., & Baldwin, T. (2010, July). Semeval-2010 task 5: Automatic keyphrase extraction from scientific articles. In *Proceedings of the 5th International Workshop on Semantic Evaluation* (pp. 21-26). Association for Computational Linguistics.
18. Isele, R., Harth, A., Umbrich, J., & Bizer, C. (2010, November). LDspider: An open-source crawling framework for the Web of Linked Data. In *Poster, International Semantic Web Conference*.
19. Lenten, L. J., & Moosa, I. A. (2003). An empirical investigation into long-term climate change in Australia. *Environmental Modelling & Software*, *18*(1), 59-70.
20. Phillips, P. C., & Perron, P. (1988). Testing for a unit root in time series regression. *Biometrika*, *75*(2), 335-346.
21. Kwiatkowski, D., Phillips, P. C., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?. *Journal of econometrics*, *54*(1), 159-178.
22. Tony Hey, Dennis Gannon, Jim Pinkelman, "The Future of Data-Intensive Science," Computer, vol. 45, no. 5, pp. 81-82, May 2012, doi:10.1109/MC.2012.181
23. Agrawal, R., Imieliński, T., & Swami, A. (1993, June). Mining association rules between sets of items in large databases. In *ACM SIGMOD Record* (Vol. 22, No. 2, pp. 207-216). ACM.
24. Data Catalog Vocabulary (DCAT). http://www.w3.org/TR/vocab-dcat/
25. Michener, W. K., & Jones, M. B. (2012). Ecoinformatics: supporting ecology as a data-intensive science. *Trends in ecology & evolution*.
26. Stodden, V. C. (2012). Data-Intensive Science: Methods for Reproducibility and Dissemination.
27. Bietz, M. J., Wiggins, A., Handel, M., & Aragon, C. (2012, February). Data-intensive collaboration in science and engineering. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work Companion* (pp. 3-4). ACM.
28. Venugopal, S., Buyya, R., & Winton, L. (2006). A Grid service broker for scheduling e-Science applications on global data Grids. *Concurrency and Computation: Practice and Experience*, *18*(6), 685-699.
29. Venugopal, S., Buyya, R., & Ramamohanarao, K. (2006). A taxonomy of data grids for distributed data sharing, management, and processing. *ACM Computing Surveys (CSUR)*, *38*(1), 3.
30. Skillicorn, D., & Talia, D. (2012). Mining large data sets on grids: Issues and prospects. *Computing and Informatics*, *21*(4), 347-362.
31. Raskin, R. G., & Pan, M. J. (2005). Knowledge representation in the semantic web for Earth and environmental terminology (SWEET). *Computers & Geosciences*, *31*(9), 1119-1125.
32. Bizer, C., & Cyganiak, R. (2006, November). D2r server-publishing relational databases on the semantic web. In *5th international Semantic Web conference* (p. 26).
33. Lang, D. T. (2007). R as a Web Client–the RCurl package. *Journal of Statistical Software, http://www. jstatsoft. org*.
34. Geser, G., Jaques, Y., Manouselis, N., Protonotarios, V., Keizer, J., & Sicilia, M. Building Blocks for a Data Infrastructure and Services to Empower Agricultural Research Communities.