The Multi-focus approach: multidisciplinary cooperations on the Web (Position paper)*

Geneviève Jomier and Claudia Bauzer Medeiros and André Santanchè

1 Problem statement

A February 2011 issue of Science [sci11] is entirely dedicated to challenges of dealing with scientific data and the data deluge (e.g., see also [Ber08, eco10]. There, 23 articles from experts working in many scientific fields exemplify open problems of dealing with data management, indexing, analysis and visualization. In all of these papers, one can find problems concerning analysis of multiple, interacting space-time scales, for which one requires scientific cooperation of groups that are distributed geographically all over the world.

Regardless of the theme, there are a few common concerns. The first one is data availability and sharing – and thus problems of consistency, security, privacy and curation. The second is the need for new data analysis and visualization mechanisms, with emphasis on data evolution through time. A third concern involves handling heterogeneity – of data and of expert domains – and hence, the issue of multiple interacting abstraction levels.

Last, but not least, most projects are concerned with how to design collaboration mechanisms on the Web and, even most important, effective means of publishing the results. The latter concern is associated with the problem of management of data associated with projects [Ada09], so that it can be reused by other groups – the very relevant issue of *scholarly communication*, one of the four main topics of [HTT09]. Scholarly communication, here, is not restricted to scientific papers – it involves the data used and produced by experiments (and by devices used in these experiments), the software designed and developed in simulations, documentation produced, etc.

This paper is concerned with discussing issues associated with the emerging paradigm of collaborative scientific environments on the Web, and on challenges facing teams with complementary expertise, who work across the Web. The emphasis is on the multiple focuses in which these groups attack a problem, and how this can be approached from a spatio-temporal database perspective.

*Work financed by CAPES-COFECUB AMIB project, and also by Brazilian financing agencies CNPq, CAPES and FAPESP, and the INCT on Web Science (CNPq 557.128/2009-9).

2 Virtual research laboratories and multi-focus work

Multidisciplinarity is breaking geographic barriers, with "implications for the location, research quality, and social stratification of science"[JWU08] – even the quality of research results needs to be studied under a new perspective.This scenario opens a wide range of new problems to be investigated. Modern multidisciplinary research is directly concerned with the challenges mentioned in this paper – scientists increasingly need to integrate results of research conducted under multiple focuses. For instance, if one considers issues raised in projects concerning climate change, there may be the need to treat multi-scale interactions that vary from the micro granularity (bacterial) to the macro one (the Earth). To exploit the evolution and interaction of such complex systems, research groups (and disciplines) must consider distinct entities of study, submitted to particular time and space dynamics. If this kind of work enables new, exciting, discoveries, it also creates a whole new set of challenges in spatio-temporal information management.

One can no longer consider data heterogeneity alone, but also the heterogeneity of processes that occur within and across granularity levels. Such heterogeneity is intrinsic to the need to combine many kinds of expertise to solve a problem, but also to involve many research groups. Though climate change is an extreme example, the same applies to other scenarios, which are increasingly common due to the fact that research laboratories are becoming virtual. The notion of Web Lab (e.g., http://icampus.mit.edu/iLabs/) is a first step – here, scientists need not physically be in a laboratory, but can access equipment remotely. Our scenario goes a step forward – a virtual laboratory is composed of several interlinked research groups, each working in a part of a problem, exchanging data and information.

This reflects two very important facts, often forgotten by computer scientists: (a) in such projects, there are distinct fields of knowledge involved (hence different collection methodologies, models and practices); and (b) the study of complex systems requires complementary ways of analyzing a problem, looking at evidence at distinct aggregation/generalization levels – a *multi-focus* approach. Since it is impossible to work at all granularities and representations at once, each group of scientists will focus on a given (sub)problem and try to understand its complex processes. The set of analyses performed under a given focus has implications on others. From now on, this paper will use the term "multi-focus" to refer to these problems, where a "focus" is a perspective of a problem, including data (and data representations), but also modeling, analysis and dynamics of the entities of interest.

In order to have a global view, scientists must integrate these partial results, necessarily by introducing simplifications or by decomposing a certain scenario. From a spatio-temporal data management point of view, one can point out the following challenges, among many others:

- How can researchers in spatio-temporal information management provide

support to research that is characterized by the need to analyze data, models, processes and events at distinct space and time scales, and represented at varying levels of detail? Here, it is not enough to consider the evolution of data – one must also provide support to the evolution of models, their interactions, and the complementary disciplines involved.

- How to keep track of events as they percolate bottom-up, top-down and across space, time and focus of interest
- How to provide adequate management of these multi-focus multi-expertise scenarios and their evolution, in our kind of virtual laboratory?

3 Some examples - multi-focus dynamics and spatio-temporal data

How do virtual multi-focus laboratories appear? Our definition of such entities is the following. A virtual multi-focus laboratory is any (virtual) research environment in which distinct groups with complementary perspectives of a research problem contribute with their expertise, data and information to construct a solution to the problem. This solution necessarily requires the participation of these groups and cannot be successfully attacked by any of these groups in isolation.

A good example of multi-focus research is multimodal transportation. At a given granularity, engineers are interested in individual vehicles, for which data are collected (e.g., average speed, itineraries). At the same time, other experts may store and query trajectories, and associate semantics to stops. At a higher level, traffic planners study trends - the individual vehicles disappear and the entities of study become clusters of vehicles and/or traffic flow [MJJV10]. Such classes are frequently correlated with urban, social or periodic spatio-temporal events (e.g., weekends, vacations, festivals). A complementary focus, still for transportation problems, comes from climate research (e.g., floods cause major traffic disturbances) or political upheavals. This can be generalized to several interacting granularity levels – from city section to country. Events at one level influence other levels (e.g., a traffic accident in a main thoroughfare may cause disturbances in a city; a long holiday has impact across countries). In spite of advances in transportation research, e.g., in moving objects, there are very few results in representation and interaction of multiple focuses.

Environmental changes, as mentioned in section 2, present a different set of challenges to spatio-temporal information management. Studies consider a hierarchy of ecological levels, from community to ecosystem, to landscape, to a whole biome. Though ecosystems are often considered closed systems for study purposes, the same does not apply to landscapes, e.g., they can include rivers that run into (or out of) boundaries[1]. A landscape contains multiple habitats, vegetation types, land uses, which are inter-related by many spatio-temporal

[1]Similar to studies in traffic in and out of a region...

relationships. And a study may focus on vegetation patches, or in insect-plant interactions. Each of the disciplines involved has its own work practices, which require analyzing data at several granularity levels; when all disciplines and data sets are put together, one is faced with a highly heterogeneous set of data and processes that vary on space and time, and for which there are no consensual storage, indexation, analysis or visualization procedures.

The example reported in [B+02], concerning riverine ecosystems, is representative of these challenges. It shows that such ecosystems involve, among others, analysis of spatio-temporal data and processes on human activities (e.g., urbanization, agricultural practices), on hydrologic properties (e.g., precipitation, flow routing), and on the environment (e.g., vegetation and aquatic fauna). This, in turn, requires cooperation of (at least) hydrologists, geomorphologists, social scientists and ecologists. "Collaborations involving multiple disciplines may create difficulties owing to mismatches in space and time scales, in forms of knowledge (e.g., qualitative versus quantitative), and in levels of precision and accuracy" ([B+02] pp. 1127). Again, just like the traffic example, an event at a given focus (e.g., change of agricultural practices) will impact others (e.g., sediment transport and deposition, impacting erosion, fauna and flora at a much larger scale).

Many other examples can be given of the need for (and problems of supporting) multi-focus analysis, and the present limitations. For instance, the need for interaction between bio-informatics and human health (from genes to human metabolism) is now taken for granted. Health scientists increasingly require cooperation with environmental experts (hence new kinds of space-time varying scales and focuses). In agriculture, the focus varies from sensors to satellites, analyzed under land use practices or crop strains and lifecycles. Thus, virtual multi-focus labs are becoming increasingly common. How do they relate to Web Science?

4 Some challenges - multi-focus collaboration on the Web

The previous sections concentrated on some challenges in handling spatio-temporal data for multiple focuses. This section analyzes a few problems associated with multidisciplinary multi-focus collaboration on the Web.

The examples of section 3 can occur either in a single research center or, as the problem scope grows, in a Web environment. For those examples, however, the specific scientific obstacles we discussed concern handling data and constraints in a multi-focus perspective, regardless of a Web scenario. If the cooperating groups (and data) are distributed, then one must add challenges that involve, for instance, fault tolerance, replication, data communication protocols, cooperation mechanisms, and data availability.

There are, however, research domains that require multi-focus approaches *because the Web exists*. A good example are problems involving the so-called

"digital humanities" [Kin11], especially social networks. Thanks to the ability to track social interactions on the Web and to store them as large graphs, social scientists are now getting to the point of having enough information to deal with previously intractable problems. This has launched research on huge graphs, at multiple levels of aggregation and clustering, with interactions across and within levels. Not only do these graphs evolve with time, but node aggregations often follow geographic criteria (directing us again to spatio-temporal issues, and "geo-sensitive" nodes).

King [Kin11] presents many examples of multi-focus interactions in social sciences, such as crime studies, analysis of political surveys, games and virtual worlds. Social networks are also at the heart of "team science" [SHTM08] – a research domain on the dynamics of team creation and interactions. Besides spatial and temporal criteria, team graphs also cluster nodes using social constraints – a distinct focus, that goes beyond the transformations proposed in the literature for spatio-temporal information management. Not only do we have physical distances, but also social ones. With the advent of the Web, this kind of work now requires collaboration of computer scientists, mathematicians and statisticians – all working at multiple interacting scales.

Another aspect concerns the Web and scholarly communication. Virtual laboratories produce a variety of data that must be made available. Even if each laboratory centralizes its "publications", there is still the need to design mechanisms that will find relevant information. In this sense, we can investigate at least the following Web Science challenges:

- Can Web-born multi-focus data management issues be approached using spatio-temporal database perspectives, or are there additional factors to consider?
- Can inter-laboratory scholarly communication solutions be treated as a generalization of solutions found for intra-laboratory communications?

5 Preliminary conclusions - possible solution directions

No single solution or approach can handle all the challenges posed by complex, multi-focus problems. Rather, several research directions must be explored to treat parts thereof. All these directions must be considered within a methodological scenario, which provides the background to support multi-focus research.

In ecosystem studies, experts often recommend that one must find a "common denominator" – by aggregating data in space and/or time, eliminating (or adding) constraints and modifying models. Distinct groups will then concentrate on different granularity levels, and focus on a subset thereof. The same applies to other fields – e.g., the multi-level graphs of digital humanities.

Finding a common denominator. How to organize multi-focus cooperations? Under this methodology, a subsystem becomes a new entity in a more

generic level (and vice-versa) – e.g., changing scale in space implies a change in the entities to be studied. Hence, each level requires distinct modeling, processing and analysis procedures (e.g., analytical models, probabilistic studies, simulation). Local changes may require just local analyses, but may also cause repercussions in other levels. Thus, one must be able to process the percolation of events throughout levels.

At the end, the results of the partial analyses have to be integrated in a global scenario. This poses problems such as keeping track of inter-level mappings and transformations (and thus traceability), or integration procedures that will vary from level to level, and that will have to take into account the very heterogeneous life cycles of the entities concerned.

A distinct, complementary, aspect is that, at a given level, experts will want to select the entities and constraints of interest, to be studied more thoroughly. Once the study is performed, these entities may have been modified (e.g., through analysis functions) and will need to be reintegrated at the appropriate granularity level. Besides posing again integration problems, this also introduces the need for establishing composition and decomposition strategies.

Using ontologies to harmonize concepts. Several computational approaches can be adopted within this methodological framework. Domain ontologies, for one, may be used to support communication among experts and interactions across levels and focuses. Here, we must face concerns such as ontology specification and management, ontology matching and merge – and, of course, obtaining consensus among experts. Moreover, given the multiple scales in space and time, additional research issues must be considered, such as ontology versioning (e.g., in biology, there are multiple versions of taxonomic classifications of species), or ontology curation (and thus the open problem of ontology quality). Ontologies can be combined with the use of data warehouses. The latter support scale changes by expansion or compression along dimensions. This is a powerful tool for multi-scale data storage and management, and for modeling multiple scales and dimensions, creating new entities via aggregation, slice and dice, but also via analysis functions. An open problem, however, is to keep track of events across such changes – and to integrate results of one level into another.

Adopting versions to create scenarios. Previous work of ours in traffic management, agriculture and biodiversity brought to light the limitations of present research on spatio-temporal information management, when it comes to supporting multi-focus studies. A promising (partial) approach is to take advantage of the DBV version model of [CJ90, DGJM96]. Here, versions can be constructed along many dimensions – versioning in time, but also in space and/or granularity level. A focus is the articulation of several versions. A version here beomes a dynamic (sub)state of the world that contains a consistent set of the entities (and models) of interest. Moreover, a version may refer to more than one scale, and thus a version may itself give origin to multiple versions (in time and/or space and/or granularity level and/or focus). All versions interact – and thus we need to treat the problem of inter- and intra-version consistency. This can be combined with research on multi-representation databases,

which address the problem of synthesizing heterogeneous entities, coming from different domains and representations, in a common denominator entity.

Constructing graphs to model multi-focus interactions. Still another direction involves advances in graph theory – besides the social sciences, huge multi-level graphs apply to, for instance, protein interaction networks or traffic analysis. It is not enough to store and process such graphs using the standard procedures of graph structures and algorithms, due to graph semantics, interactions across levels, and the dynamics of graph evolution. Present graph-based algorithms are not able to cope with these issues.

References

[Ada09] Adaguc. Atmospheric data access for the geospatial user community, http://adaguc.knmi.nl/, 2009.

[B+02] L. E. Benda et al. How to Avoid Train Wrecks When Using Science in Environmental Problem Solving. *Bioscience*, 52(12):1127–1136, 2002.

[Ber08] F. Berman. Got data?: a guide to data preservation in the information age. *Commun. ACM*, 51:50–56, December 2008.

[CJ90] W. Cellary and G. Jomier. Consistency of Versions in Object-Oriented Databases. In *Proc. 16th VLDB*, pages 432–441, 1990.

[DGJM96] A. Doucet, S. Gancarski, G. Jomier, and S. Monties. Integrity Constraints in Multiversion Databases. In *Proc. BNCOD*, pages 56–73, 1996.

[eco10] Special issue - the data deluge, February 2010.

[HTT09] T. Hey, S. Tansley, and K. Tolle, editors. *The Fourth Paradigm: data-intensive scientific discovery*. Microsoft Research, 2009.

[JWU08] B. F. Jones, S. Wuchty, and B. Uzzi. Multi-University Research Teams: Shifting Impact, Geography, and Stratification in Science. *Science*, 322, 2008.

[Kin11] G. King. Ensuring the Data-Rich Future of the Social Sciences. *Science*, 331(6018):719–721, 2011.

[MJJV10] C. B. Medeiros, M. Joliveau, G. Jomier, and F. Vuyst. Managing sensor traffic data and forecasting unusual behaviour propagation. *Geoinformatica*, 14(3), 2010.

[sci11] Dealing with Data - Challenges and Opportunities - Special Issue. *Science*, 331, February 2011. http://www.sciencemag.org/site/special/data/.

[SHTM08] D. Stokols, K. L. Hall, B. K. Taylor, and R. P. Moser. The Science of Team Science - Overview of the Field. *American Journal of Preventive Medicine*, 35, 2008.