Multi-Scale Data Integration Challenges in the Observational Science Data Space

Herausforderungen der Datenintegration wissenschaftlicher Daten bei variabler Größenordnung

Laure Berti-Équille, Institut de Recherche pour le Développement, Aix-en-Provence, France

Summary In Europe, more than one thousand of laboratories intensively collect data to measure various properties of the Earth. Scientists observe environmental conditions, ecosystems and biological species. The ability to understand complex phenomena (e. g., global warming) and predict trends from spatio-temporal data becomes a major issue in observational science. However, theoretical and technical advances in multi-scale data integration are necessary to achieve this. This paper will present some challenging research directions for integrating such massive multi-scale scientific data. ▶▶▶ **Zusammenfassung** In Europa sammeln mehr als tausend Laboratorien Daten, um verschiedene Eigenschaften der Erde zu messen. Die Fähigkeit, komplexe Phänomene (z. B. die globale Erwärmung) zu verstehen und Trends aus räumlich-zeitlichen Daten zu prognostizieren ist ein wichtiges Thema in der beobachtenden Wissenschaft. Allerdings sind theoretische und technische Fortschritte in der Integration von Daten unterschiedlichster Größenordnung (Multi-Scale Data Integration) notwendig, um dies zu erreichen. Dieser Beitrag stellt einige Forschungsherausforderungen in diesem Kontext vor.

Keywords H.2.8 [Information Systems: Database Management: Database Applications]; scientific databases, new trends in data integration, geoscience data, multi-scale data integration, spatio-temporal data ▶▶▶ **Schlagwörter** Neue Trends in der Datenintegration, geowissenschaftliche Daten, multi-skale Datenintegration, räumlich-zeitliche Daten

1 Observational Science Data

To understand the environmental crisis of our planet, scientists need to systematically collect all the observable features from the Environment and the Earth spheres and examine the deep interactions between the climate change, environmental degradation, urbanization, energy production, etc. In Europe, more than one thousand of scientific laboratories from governmental organizations and private industry collect data using various sensors on board of research vessels, submarines, fixed and drifting platforms, radio-sondes, airplanes, satellites, land-based environmental stations to measure physical, geophysical, geological, biological and chemical parameters of the Earth, environmental conditions, biological species, and ecosystems.

Large-scale efforts are underway at regional and global levels. For instance, the European *INSPIRE* Directive (2007/2/EC)[1], the world-wide GEOSS (*Global Earth Observing System of Systems*)[2] common infrastructure and its European contribution, GMES (*Global Monitoring for Environment and Security*)[3] provide large catalogues of Earth observation and environmental data sets with dir-

[1] INSPIRE Directive: http://inspire.jrc.ec.europa.eu/
[2] GEOSS: http://www.earthobservations.org/
[3] GMES: http://www.gmes.info/

ect access to online services and digital data repositories. FP7 e-infrastructures such as *GENESI-DR*[4] – *Ground European Network for Earth Science Interoperations – Earth Science Digital Repositories* (2008–2009) and *GENESI-DEC – Digital Earth Community* (2010–2012) provide access to heterogeneous standardized data (airborne, *in situ*, satellite), service discovery and on-demand processing capabilities.

While these emerging infrastructures have a clear motivation (e. g., understanding global change and environmental variability) and successful results in providing instruments and systems that now generate millions of Environment and Earth system-related data sets, it is less obvious how effectively and efficiently scientists from various disciplines can collaborate through a common process of scientific knowledge discovery. This is mainly due to extremely complex data integration issues involving analytics and scaling methods to handle multi-scale spatio-temporal data sets.

On the other side of the user spectrum, more and more individuals and communities with unprecedented high attention paid to climate change and environmental issues actively collaborate in providing and discussing online volunteering information (e. g., on Google Earth and Wikimapia). Data created by amateur citizens can provide an interesting alternative to traditional authoritative information from mapping agencies but the data quality of this new source is a major concern, since volunteered information carries none of the assurances that lead to trust in officially created data. However, once data are available, it seems essential for both user communities to be assisted in each task towards collaborative knowledge discovery: from data quality assessment and data integration to exploration, interpretation and corroboration of data and results.

Beyond the accessibility and interoperability of thousands of large open science data repositories, there is still a lot of work to be done in terms of efficient, on-demand and fine-grained data integration, data quality control, collaborative data exploration and multi-data set analysis.

In the context of environmental monitoring, remote sensing images have become essential tools, as shown by the large number of international programs for the acquisition and provision of such observation data (e. g., CNES Orfeo[5], GMES[3] of ESA, NASA ESDIS[6]). Internationally, NASA (*Earth Science Technology Office*) has funded many projects around the modeling of environmental phenomena through data integration and data mining techniques. In 2004, the "*Mining Massive Earth Science Data Sets for Climate and Weather Forecast Models*" was based on the integration, summarization and mining of multi-source observation data. *National Science Foundation* also supported projects on this theme such as "*Spatio-Temporal Data Mining for Global Eco-Scale Climatic Data*" in 2007–2010. The project "*Transform Mechanism among Observed Data, Geo-Spatial Information and knowledge*" (2005–2010) uses the techniques of neural networks and SVM ("Support Vector Machine") to extract and integrate spatio-temporal information from weather data and satellite imagery.

The special nature of Earth observation data renders traditional data integration techniques inadequate. We can summarize the main technical issues as follows:

Spatio-temporal nature of observational data. The complex geometry of the observed spatial objects (or phenomena) combined with their temporal nature and variability (e. g., changing shape, reflectance, instantaneous or durable) complexify the problem of entity identification and resolution; spatio-temporal relationships between objects may be detectable only at a specific scale (in time, space or satellite image resolution) but traditional data integration techniques lack advanced analytic methods to take advantage of spatial and temporal autocorrelation from one scale to another. Furthermore, they are not able to capture long-range dependencies of non-stationary spatio-temporal objects and take advantage of them for multi-scale data mapping.

Scalability and high-dimensionality. The size of Earth science data sets can be very large, especially for high-resolution vegetation data. For example, for each time instance, one satellite image covers $2.5° \times 2.5°$ and provides hundreds of indicators characterizing about 10 000 locations of the globe; for an image resolution of $250\,m \times 250\,m$, same, similar or other indicators can be computed for about 10 billion of locations, and similarly for $50\,m \times 50\,m$ resolution, indicators are available for about 250 billion of locations; all of these indicators have to be integrated and merged in some cases to avoid conflicts over long time series (e. g., LANDSAT[7] images cover $185\,km \times 185\,km$ since 1972).

Semantic and analytic heterogeneity. The recent interest shown by NASA (which created the SWEET ontology – Semantic Web for Earth and Environment Terminology [14] and GEOSS – *Global Earth Observation System of Systems*) for ontologies opens new perspectives of their application to Earth observation science (see, e. g., the work of extending the SWEET ontology to hydrogeology and ecology [13; 16]). In the Geoscience domain, the stage of maturity for data integration is comparable to the first approaches in biomedical sciences [1; 2; 10]. The development of new methods and new tools to better suit the needs of scientists for integrating various data sets from different disciplines (biology, botany, ecology, geography, hydrology, seismography, remote sensing, etc.) requires the development of domain-specific ontologies to describe Earth, environmental and biodiversity indicators.

4 GENESI-DR: www.genesi-dec.eu/
5 CNES Orfeo http://smsc.cnes.fr/PLEIADES/A_prog_accomp.htm
6 NASA ESDIS: http://earthdata.nasa.gov/about-eosdis/esdis-project
7 Landsat Imagery: http://glcf.umd.edu/data/landsat/

In this context, raw data from sensors (e. g., for humidity, temperature, etc.), data "collected from the field" (e. g., by botanists, geologists, etc.), and processed data derived from advanced analytic or satellite imagery techniques (after segmentation and classification) have to be integrated altogether although they don't have neither the same abstraction level nor the same degree of processing.

2 Illustrative Example of Scientific Knowledge Discovery Requiring Multi-Scale Data Integration

To illustrate the inadequacy of state-of-the art solutions for scientific data integration, let us consider two examples in environmental and health sciences:

1) ***Predicting long-term environmental impacts of a region***

- *Environmental data:* Given a change in demographics and socio-economic profiles of a region (e. g., implantation of a new industry, excessive draining out of ground water, building of a highway into a fragile hilly environment, or deforestation for housing infrastructure development), scientists want to predict long-term environmental impacts on this particular region. To design possible scenarios, various sets of geo-referenced data are required;
- *Geological and hydrological data:* From the analysis of the relief, surface/ground water, soil, and vegetation/crop input data, a series of surface/ground water and soil data have to be collected to characterize the hydro-geological conditions of the region, showing existing spatial as well as temporal variations. For example, with establishment of an industrial zone, the water balance situation would change, depending upon the water needs of the industry and the increased population, and the extent of deforestation done for this purpose;
- *Demographic and socio-economic data:* The industry could create new employment; some people may shift from the agriculture industry to jobs. This region could attract people from far away regions with less economic opportunities. There would be a need for new infrastructures to support the sudden changes (e. g., new houses, roads, utilities, electric power station, etc.);
- *Wildlife data:* The wildlife in this region would get impacted, and there would be a need for defining protected areas. Impact on wildlife may impact the tourism potential of the area, which would further impact the economic factors in the region;
- *Air, water and soil pollution data:* This is linked to the nature and magnitude of the pollutants dispersed by the industry (based on the nature of chemicals/gases released and their prior treatment, the source and the sink locations), growth of population in that region, extent of deforestation, and the shift to growing of cash crops which may affect the land fertility. The overall impact of the pollution is linked to the tourism potential of the area, and to the health conditions of the local inhabitants. There would be an expected change in all these aspects over time, and the long-term demographic, economic and ecological scenario could be quite different from the short-term ones.
- *Hazard data:* Location of sources causing industrial pollution/contamination risk, nature of pollutant/contaminant and probability of leakage/accident, because of natural or manmade factors, "footprint" or the concentration map for such scenarios can be collected or predicted based on the historic data of past similar events or on weather data and fluid dynamics modeling.
- *Meteorological data:* Based on weather and climatic data history, the dispersion of the pollutant/contaminant originating from a source at a certain height, and spreading in the direction of expected wind direction has to be modeled accounting for the uncertainties associated with the hazardous material release pattern, wind direction, magnitude, etc.
- *Remote sensing data:* In order to understand how deforestation occurred in the region of interest during the last decade for example, data and studies may be derived from Landsat TM and ETM+ imagery at a resolution of 28.5 meters daily over several time frames. Similarly, other resolutions, satellite images and vegetation indices (e. g., MODIS[8] 10° latitude ×10° longitude, every 16 days, with resolution 250 m) may be used as well.

  Many other data sets (e. g., describing the seismic activity of the region, public health problem outbreaks, epidemics, traffic flows, energy production/consumption, etc.) with different data quality levels and data types can be added along the way. These data sets can be provided by observatories, research institutes, governmental or private organizations, HMOs, insurance companies or even amateur citizens. Opinions of scientists and experts from various disciplines are essential in developing a set of meaningful scenarios for predicting long-term environmental impacts. The exchange of such annotations associated with the data, analysis results, and by-product recommendations by experts is also critical for the interpretation and validation of results and predictive scenarios. Beyond the scientific communities and their needs for advanced integration techniques for multi-scale spatio-temporal data, the average citizens could gain valuable environmental awareness if they can visualize, easily understand, follow and discuss the data related to all the environmental aspects of the region where they live, and integrate new volunteered information on-demand.

2) ***Climate-dependent epidemiology***

In the context of epidemiologic surveillance (e. g., vector-borne diseases such as malaria, schistosomiasis, dengue), various data types have to be continuously gathered,

[8] MODIS: http://modis.gsfc.nasa.gov/

analyzed, and interpreted to understand disease precursors, evolution and propagation. This surveillance is essential for monitoring endemic transmission and for early recognition of impending epidemics. Moreover, to understand how climate might affect the incidence of vector-borne diseases, one must first examine the life cycles of the diseases and the environmental parameters associated with each stage. The transmission of vector-borne diseases to human populations depends upon the attributes and requirements of the pathologic agent (e. g., a virus, protozoa, bacteria, or worm), the vector (usually arthropods such as ticks or mosquitoes), and the human host. Typically, various data with different time scales have to be collected, integrated and cross-analyzed: epidemiologic data (e. g., patient age, sex, blood serotype), entomological indices (e. g., number of infected houses), environmental data (e. g., forest coverage by satellite imagery), vector surveillance data (e. g., mosquito density), and meteorological data (e. g., precipitation, humidity, temperature, wind) with time stamps and GPS locations. Some data is captured every month over more than fifty years time frames, some other every 5 seconds or 10 minutes over shorter periods of time. In this particular context, temporal multi-scale data integration is extremely complex and error-prone since data mappings have to be operated at different scales and levels of granularity in time and space.

These examples illustrate some of the new challenges of multi-scale data integration that lie in connecting multiple data sets that are diversely designed, with various spatial, temporal, semantic and structural scales that require not only schema-based and instance-based integration, but also advanced analytic mappings as well as ontological alignments.

3 Multi-Scale Data Integration Challenges

The spatio-temporal and multi-scale nature of observational science data sets requires the development of new methods for exploring and integrating such massive and complex data. These solutions must adapt to changes in scientific measurement and imagery processing tools, while allowing the detection of breaks or trends in the studied phenomena that may have a direct impact on data transformation and integration.

In addition to the distributed and multi-scale dimensions of scientific data, its dynamic dimension requires online data integration and active data cleaning tools. Technically, observation science data includes "static" structured data and a combination of streaming data at various time/space frames, frequencies and resolution granularities. On the one hand, current integration technology based on adapters to support streaming data integration could be used since they are relatively straightforward as long as the source data can be correctly represented in the data model, albeit involving some manual work and inflexibility [15]. Unlike in the traditional setting, some observational data is usually not stored in advance and has to be pulled and processed by imagery, mathematical modeling or simulation primitives before the integration and only when needed. Consequently, several input sources (and processing workflows) can be unpredictable and unreliable; delays, losses, disorders will have to be dealt with during the integration of these data sources. On the other hand, adapters should also be very light-weight so as not to become a bottleneck by slowing down the rate at which data can get in and out of the integration pipeline.

Integrating observational science data sources involves:

- ***A semantic component***, when schemas of inputs from sources across different disciplines have to be mapped to one another; this component is supported by conventional approaches [3; 7; 9].
- ***An analytic component*** dealing with the scaling theory issues [6; 17] and the heterogeneity of abstraction and aggregation levels between the data sources to integrate (some providing raw data, others providing pre-processed, analyzed and derived information).

Moreover, solutions for integrating multi-scale spatio-temporal Geoscience data must consider the changing quality and reliability of observational data, either because of the difficulties and uncertainties of sensors, satellite instrumentation, and measurement conditions, either because of specific policies related to each database or participating agencies [11].

The ultimate goal of an integrative project in observational science is to better understand a complex spatio-temporal phenomenon from multiple data sets generated by various experimental or instrumental protocols and simulations from different disciplines and describing different granularity levels, global or partial views in time or space of the same phenomenon.

In this context, the main challenges of observational science data integration lie in addressing the following issues:

- ***Temporal, spatial, structural, semantic or analytic dependencies.*** For example, two data sets of the same phenomenon can be gathered within two overlapping temporal windows using different instrumentation devices and data collection frequencies; one data set may contain measurements taken for one single GPS location and the other data set may cover a wider area including the former location. Two data sources may also be dependent [4]. Another example from biomedical data, would be to better understand complex dynamic mechanisms involved in the treatment and the evolution of a pathology: from gene interactions and intracellular signaling, to cell-level movements and interactions, to vessel branching and capillary network formation, to tissue level characteristics, to organ system response to the treatment [1; 2]. In these examples, some data may be partially redundant and overlapping or it may present the same phenomena from different perspectives, which may not be easily and semantically re-unified.

- ***Different levels of data granularity or data abstraction from raw measurement data to processed data and derived statistics***. Consider a data set that gives the exact precipitation value per hour at a given GPS location and another data set that provides the average precipitation value per month over a wider area embedding the previous location. Depending on the analytic problem the scientist may try to solve, the "standard" solution will not necessarily be to choose one or the other value, but also to extrapolate the precipitation values on each location at a finer grain from the average value of the embedding region and check for consistency. This scenario will inevitably lead to conflicting data values and involve complex analytic simulations to select the best strategy for data fusion, thus extending the current approaches [5].
- ***Various data interpretations or usages depending on the disciplines***. For example, to understand the evolution of biodiversity in a particular region, many data sets from various fields and disciplines are required, such as climate and meteorological data, botanical surveys of the species living in the area, geological and geochemical measurements, and regional environmental specificities. Since these data sets describe various parameters of the same region, they have to be mapped and integrated. This requires multi-disciplinary expertise, multiple ontologies and their corresponding accurate alignment. In this typical case, it is not sufficient to collect the parameters from various disciplines as if they were independent. Although they may not overlap (i. e., the data values are neither identical nor similar from one source to another), they are correlated and dependent. And we claim that ***analytic mapping*** and discovery of dependencies is actually one of the main challenges of observational science data integration.
- ***Quality heterogeneity of spatio-temporal data***. Errors and uncertainty are facts of life in all data sources. It is even worse when the data set is obtained from multi-scale spatio-temporal data sources. Typically, the data quality problems may be present as unexpected errors, data conflicts, outliers, duplicates, inconsistencies, and so on. The conflicts reconciliation strategies proposed in the non-spatial data integration systems are not readily applicable for spatio-temporal data sets. This is mainly because some spatial relations and temporal relations can be misrepresented and not captured.
- ***Scaling issues.*** Scale effect in space and time is a challenging research issue in data integration. Scale, in terms of spatial resolution or temporal granularity, can have direct impacts on the kinds and/or strength of relationships that can be identified in the data sets and used for their integration. For example, in climatology, currently coupled atmosphere-ocean general circulation models (GCMs) generate projections (i. e., simulations) on a too coarse scale to obtain information at local-scale level. These models solve the principal physics equations of the dynamics of the atmosphere and of the oceans together with their interactions over the globe. GCMs allow scientist to simulate climate variables and to study the mechanisms of the present, past, and future to evaluate the potential impacts of climate changes on economy, agriculture, and ecology in the next decades [16]. Such impact studies require climate simulations at high spatial resolution (small scale), ranging from a few kilometers down to station locations. In particular, precipitation, which is of major importance in agriculture, vegetation, and flood risk assessment, has a strong spatial variability. However, the spatial resolution at which GCMs operate (about 200×200 km) is typically too low to capture such spatial variability. Other reasons why GCMs struggle to reproduce precipitation are related to the features of the distribution of precipitation, namely, boundedness at zero, non-normality, and the presence of extreme values at local scale with a potential destructive power. In this context, downscaling techniques have been developed to bridge the gap between large- and small-scale variables. They are commonly used to infer from climate simulations of models to local-scale projections. There are two different approaches to downscaling: the dynamical approach consists in refining GCMs over a higher-resolution grid. These refined GCMs, called regional climate models (RCMs), operate at a resolution down to about 10 km. RCMs have a high computational cost and thus are often limited in their uses to restricted regions and periods of time. However, some processes (e. g., hydrological processes) typically occur on finer scales than those provided by GCM outputs. Hydrological regional models are in general driven with downscaled data, normally temperature and precipitation data. But precipitation is difficult to model and to downscale, mostly due to its high spatial and temporal variability, missing values and its nonlinear nature. By taking advantage of extreme value theory and recent developments in Geostatistics and Bayesian analysis (e. g., in [6; 11] with statistical downscaling of extremes, i. e., modeling of the relationships between extremes recorded at different spatial scales), the challenge is to develop innovative downscaling schemes to be incorporated into multi-scale data integration for Geoscience.

4 Concluding Remarks

The need to investigate both "spatial" and "temporal" relations at the same time complicates the data integration and mapping tasks even further. A crucial challenge in spatio-temporal multi-scale data integration is the exploration of efficient up- and downscaling methods scalable to large amounts of scientific data and robust to the complexity of spatio-temporal data types, data representation, and spatial data structure.

Spatial and temporal relationships exist among spatial entities at various levels (scales). One of the purposes of spatio-temporal multi-scale data integration is to capture such relationships and use them for mapping. The spatial relations, both metric (e. g., distance) and non-metric (e. g., as topology, directions, shape, etc.), and temporal relations (e. g., as before or after) may be explicit or implicit in scientific databases. In both cases, such relationships are information bearing and therefore need to be considered by new analytic integration techniques.

References

[1] J. Bleiholder, Z. Lacroix, H. Murthy, F. Naumann, K. Raschid, and M-E. Vidal. BioFast: Challenges in Exploring Linked Life Science Sources. In: *SIGMOD Record*, 33(2), 2004.

[2] S. Cohen-Boulakia, and U. Leser. Next Generation Data Integration for the Life Sciences. In: *Int'l Conf. on Data Engineering (ICDE)*, Seminar, 2011.

[3] A. Doan and A. Y. Halevy. Semantic integration research in the database community: A brief survey. In: *AI Magazine*, 26(1):83–94, 2005.

[4] X. L. Dong, L. Berti-Équille, and D. Srivastava. Integrating conflicting data: the role of source dependence. In: *Proc. of the 35th Int'l Conf. on Very Large Data Bases (VLDB)*, Vol. 2, Issue 1, pages 550–561, 2009.

[5] X. L. Dong and F. Naumann. Data fusion-Resolving data conflicts for integration. In: *35th Int'l Conf. on Very Large Data Bases (VLDB)*, Tutorial, 2009.

[6] M. Ghil, O. Yiou, S. Hallegatte, B. D. Malamud, P. Naveau, A. Soloviev, P. Friederichs, V. Keilis-Borok, D. Kondrashov, V. Kossobokov, O. Mestre, C. Nicolis, H. W. Rust, P. Shebalin, M. Vrac, A. Witt, and I. Zaliapin. Extreme events: dynamics, statistics and prediction. In: *Nonlinear Processes in Geophysics*, 18, 295–350, 2011.

[7] A. Y. Halevy. Information Integration. In: *Encyclopedia of Database Systems*, 1490–1496, Spirnger 2009.

[8] Intergovernmental Panel on Climate Change (IPCC). *Climate Change 2007: Impacts, Adaptation and Vulnerability*. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, edited by M. L. Parry et al., Cambridge Univ. Press, Cambridge, U. K., 2007.

[9] Z. G. Ives. Data Integration and Exchange for Scientific Collaboration. In: *Proc. of the 6th Int'l Workshop on Data Integration in the Life Sciences (DILS)*, pages 1–4, 2009.

[10] M. Lenzerini. Data Integration: A Theoretical Perspective. In: *21st ACM SIGACT-SIGMOD-SIGART Symp. on Principles of Database Systems (PODS'02)*, Tutorial, pp. 233–246, 2002.

[11] C. Lynnes, E. Olsen, P. Fox, B. Vollmer, R. Wolfe, and S. Samadi. A quality screening service for remote sensing data. In: *Proc. of the 19th ACM Int'l Symp. on High Performance Distributed Computing* (HPDC '10), pages 554–559, 2010.

[12] D. Maraun, F. Wetterhall, A. M. Ireson, R. E. Chandler, E. J. Kendon, M. Widmann, S. Brienen, H. W. Rust, T. Sauter, M. Themeßl, V. K. C. Venema, K. P. Chun, C. M. Goodess, R. G. Jones, C. Onof, M. Vrac, and I. Thiele-Eich. Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. In: *Review of Geophysics*, 48, 2010.

[13] J. S. Madin, D. Bowers, M. P. Schildhauer, and M. B. Jones. Advancing ecological research with ontologies. In: *Trends in Ecology & Evolution*, 23(3):159–168, 2008.

[14] R. G. Raskin and M. J. Pan. Knowledge representation in the Semantic Web for Earth and Environmental Terminology (SWEET). In: *Computers & Geosciences*, 31, pp. 1119–1125, 2005.

[15] N. Tatbul. Streaming Data Integration: Challenges and Opportunities. In: *Proceedings of the IEEE 26th Int'l Conf. on Data Engineering (ICDE), Int'l Workshop on New Trends in Information Integration (NTII'10)*, pp. 155–158, 2010.

[16] A. Tripathi and H. A. Babaie. Developing a modular hydrogeology ontology by extending the SWEET upper-level ontologies. In: *Computers & Geosciences*, 34, pp. 1022–1033, 2008.

[17] P. Willems and M. Vrac. Statistical precipitation downscaling for small-scale hydrological impact investigations of climate change. In: *Journal of Hydrology*, 402, 193–205, 2011.

Received: December 6, 2011, accepted: March 20, 2012

Dr. Laure Berti-Équille received a PhD degree in computer science in 1999 from University of Toulon, France. In 1999–2000 she worked as an Assistant Professor at University of Avignon. In 2000 she joint IRISA as a permanent Associate Professor of University of Rennes 1. In 2007–2009 she was a visiting researcher at AT&T Labs Research NJ USA. Since 2011, she joint IRD as a Senior Research Scientist. Her research interests include data quality management and data mining.

Address: CEREGE IRD, Europôle Méditerranéen de l'Arbois, B. P. 80, F-13545 Aix-En-Provence cedex 04, France, e-mail: laure.berti@ird.fr