DATA INTEROPERABILITY

Pasquale Pagano, Leonardo Candela, and Donatella Castelli

Networked Multimedia Information Systems Laboratory (NeMIS), Istituto di Scienza e Tecnologie dell' Informazione (ISTI)"Alessandro Faedo", Italian National Research Council (CNR), Via G. Moruzzi 1, 56124 Pisa, Italy
Email: pasquale.pagano@isti.cnr.it; *leonardo.candela@isti.cnr.it; donatella.castelli@isti.cnr.it*

1 STATE OF THE ART

In the context of scientific investigations, data have acquired an ever growing leading role while their large scale, cross-community and cross-domain sharing have concurred to identify new investigation paradigms (Hey, Tansley, & Tolle, 2009). Unfortunately, data interoperability – a mandatory prerequisite for achieving the above scenarios – is still a difficult open research challenge. Both the "data" and "interoperability" concepts are difficult to be fully perceived and actually lead to different perceptions in diverse communities. This problem is further amplified when considered in the context of (global) research data infrastructures that are expected to serve a plethora of communities of practice (Lave & Wenger, 1991) potentially involved in very diverse application scenarios, each characterised by a specific sharing problem.

The term "data" is on its own very common yet difficult to define because it may take many forms, both in the digital and in the real world. Moreover, the act of recognising or understanding that "something" – e.g., observations, statistics, artefacts, records – constitutes data is an intellectual activity that is usually driven by a certain goal. The term "research data" – here used to refer to the kind of data a (global) research data infrastructure has to deal with – adds another factor of difficulty. Such kind of data is collected for many purposes, via different approaches, and very often it is difficult to interpret once exploited in contexts other than its initial one (Borgman, 2010; Borgman, 2011). Research data range from traditional research outputs, mainly papers and data, to living reports (Candela, Castelli, Pagano, & Simi, 2005; Candela, et al., 2007), executable research papers (Van Gorp & Mazanek, 2011; Nowakowski, et al., 2011), and scientific workflows (De Roure, Goble, & Stevens, 2009). Very often these data fall into the category of "big data" (Stapleton, 2011), i.e., data characterised by (i) volume, i.e., their dimension in terms of bytes is huge; (ii) velocity, i.e., their speed requirements for collecting, processing, and using is demanding; and (iii) variety, i.e., their heterogeneity in terms of data types to be managed and data sources to be merged is high.

The term "interoperability", although heavily used to describe a core class of problems in many systems and application scenarios, does not yet have a clear definition shared by the overall community. The IEEE Glossary – in 1991 – defined interoperability as "the ability of two or more systems or components to exchange information and to use the information that has been exchanged" (Geraci, 1991). Over the years, other definitions and characterisations have been proposed each focusing on certain aspects, e.g., syntactic versus semantic (Wegner, 1996; Heiler, 1995; Park & Ram, 2004). Asuncion and van Sinderen (2010) discussed "pragmatic interoperability", i.e., the interoperability dealing with mutual understanding in the use of data between collaborating systems. Recently, comprehensive frameworks have been proposed to capture the many facets of interoperability (Candela, Castelli, & Thanos, 2010; European Commission, 2004). These frameworks bring in interoperability facets that are usually overlooked and contribute to show that the myth that relegates interoperability to a merely technical issue is false. Interoperability actually is a problem affecting the interaction of entities at various levels including: (i) Organisational level, i.e., business goals and processes of the institution operating every single entity involved in the interoperability scenario; (ii) Semantic level, i.e., meaning of the exchanged digital resource – mainly data or part of it in the case of data interoperability – including their contextual information; and (iii) Technical level, i.e., the heterogeneity in technology supporting the operation of every single entity involved in the interoperability scenario including the communication channel and the information exchanged through it.

All too commonly, data interoperability, data integration, and data exchange are confused, possibly because they share some commonalities in terms of issues and goals. In data integration the goal is to synthesize data from different data sources – usually independent of each other – into a unified "view" according to a "global" schema (Lenzerini, 2002; Halevy, Rajaraman, & Ordille, 2006). In data exchange the goal is to take data from a given data source and transfer them to a target data source such that it "reflects" the given source data as

accurately as possible (Kolaitis, 2005). Although both data integration and data exchange problems – described as the "oldest database problems" – have been largely investigated, they are still considered extremely hard tasks (Anthes, 2010).

Implementing data interoperability requires realising data integration and data exchange as well as enabling effective use of the data that become available. Each of these three tasks involves some type of schema matching (Rahm & Bernstein, 2001) and schema mapping (Kolaitis, 2005), two approaches that capture the relationships between the source and the target data, i.e., the data resulting from the combination of the source data to serve the needs of the entity that is going to consume them.

Sometimes scientific literature distinguishes between issues related to 'data interoperability' and to 'metadata interoperability'. In most cases, this distinction is unjustified. Metadata is usually defined as 'data about data' or 'data that describe resources, digital or non-digital' (Haslhofer & Klas, 2010). Because of this, there is no piece of information that is only metadata; every piece of information is primarily data per se and assumes the qualification of metadata when considered in a context where this information is associated with another piece of information with the goal of capturing some of its characteristics. With respect to interoperability, although metadata interoperability and data interoperability have, for a long time, been considered two different problems, their solutions and approaches share many commonalities, e.g., both might necessitate building mappings, and this is among the tasks that have a considerable cost.

The current approaches and solutions to data interoperability can be classified as: 'Agreement-based' approaches and 'Mediator-based' approaches. Usually, interoperability scenarios are complex and require the combination of multiple solutions to be resolved. Even in this case, the constituent solutions are either agreement-based or mediator-based. In some cases agreement-based and mediator-based approaches blend into each other. Agreement-based approaches consist of agreement on a set of principles that enables the achievement of a limited amount of homogeneity among heterogeneous entities. Standards belong to this category. The value of standards is clearly demonstrable. However, in the majority of cases their adoption by an organization comes after a careful evaluation of pros and cons. Standards often end up being complex combinations of features reflecting the interests of many disparate parties. Thus it is not always simple to implement them. Moreover, by nature they infringe on the autonomy of the entities adopting them. This means that organizations have to agree on changing their behaviour in favour of improved interoperability. Examples of large-scale agreement-based approaches and initiatives for promoting data interoperability include Linked Data (Heath & Bizer, 2011), INSPIRE (European Parliament, Council, 2007), SDMX (SDMX Initiative), and OAI-ORE (Lagoze & Van de Sompel, 2008).

Mediator-based approaches have been proposed to resolve scenarios where there is a need to guarantee a high level of autonomy among the partaking entities. These approaches isolate the interoperability machinery and implement it in components specifically conceived to link the entities involved in the scenario. These solutions have initially been conceived in the Information Systems domain (Wiederhold & Genesereth, 1997) and are nowadays realised in many ways and exploited in many diverse scenarios. All the approaches based on some sort of mappings, e.g., schema mapping and instance transformation, belong to this class of solutions.

2 TEN YEAR VISION

In ten years from now, it is expected that a unified information space – which may be a virtual information space built by aggregating a number of information spaces that have been built by aggregating content from multiple information sources – will be in place thanks to a data infrastructure. This information space will give seamless access to heterogeneous data originally scattered across a number of independent data sources. However, this information space is not based on a "tight" data homogenization approach; instead, it provides for (a) heterogeneous data sources co-existence and (b) on-demand and flexible integration and exchange of data. The data infrastructure will provide basic functionalities for all data sources, regardless of how they interoperate, as well as a rich array of sophisticated functionalities for (a) enhancing the level of interoperability among any target set of data sources with respect to the needs arising in a specific application context and (b) sharing resources – e.g., data mappings and transformation facilities – to reach a certain level of interoperability among data sources whose characteristics are known. Thus, the data infrastructure supporting the unified information space endorses the idea of pay-as-you-go; it guarantees that some services promoting interoperability are immediately offered at no setup cost while the overall service is improved as more investment is made into creating resources tailored to make a (set of) data source(s) interoperable at a higher level.

3 CURRENT CHALLENGES

As a highly challenging and multifaceted task, data interoperability subsumes a lot of challenges and research topics including: lack of a common problem definition, coping with variety, enabling data reuse, agreeing on common standards, and developing comprehensive approaches.

The lack of a common problem definition might seem a trivial aspect. On the contrary, it is a fundamental issue that prevents the overall community from working in synergy towards the identification of proper strategies and solutions. As already mentioned, there is neither a definition of interoperability nor one of data and data interoperability that is shared across different communities and domains. However, different communities have faced and will face data management and data interoperability problems in the context of their application scenarios. Very often these communities have no specific expertise in data interoperability and will follow a pragmatic approach oriented to resolving the specific issue they are confronted with. Very often they develop from scratch an ad-hoc solution while having no, or very limited, knowledge of approaches and solutions developed by other communities in similar cases. All this is a result of the lack of a common framework that can be used to describe the interoperability problem in all its facets in a structured and unified manner – well beyond the technical interoperability that although fundamental is only a part of the problem. Once available, such a framework could be used to describe interoperability approaches and solutions in a systematic way in order to ensure that they are discovered thus avoiding a scenario where these valuable resources remain confined to the domain in which they have been developed.

Coping with variety is a very broad yet characterizing aspect of interoperability. Variety is a characteristic spanning the entire spectrum of data features when they come from different and independent data sources. In a (global) research data infrastructure, data to be managed might be heterogeneous with respect to their type, accuracy, size, semantic, etc. Some of these variety aspects are objective or application-agnostic, i.e., they exist independently of the characteristics of the interoperability scenario, while others are subjective or application-specific, i.e., their existence depends on the needs of the specific interoperability scenario. The boundary between objective and subjective aspects is a parameter difficult to estimate; in many cases it is application-specific. To make it possible for every "user" of data to decide which are the data variety aspects owned by given data that might be tolerated and which are those to be removed, a common approach is to enrich the data with others' data, capturing data variety, e.g., data provenance (Moreau, 2010) as well as data annotations and metadata aiming at characterising data quality aspects (Batini & Scannapieco, 2006). Unfortunately, there is neither a standard universally accepted for these data nor a widely accepted approach for dealing with diverse materialisations of such "additional" data. This calls for approaches for data provenance / data quality interoperability.

Enabling data reuse is an aspect that characterises data interoperability with respect to similar problems, e.g., data exchange. In some contexts data have value if and only if they can be re-used. In order to make it possible for an entity to actually re-use data that have been collected or produced by a different entity, it is fundamental that a rich set of contextual information about such data be made available. Therefore, open research problems are (a) the characteristics this set of contextual information should capture, (b) the format this information should be represented in, and (c) the manner this information should be communicated.

The development and wide adoption of common standards is extremely difficult despite the fact that they represent the most effective and powerful tool to deal with interoperability. There is a broad and powerful array of forces driving evolution and processes in communities of practice. Standards are difficult to agree on and very often end up being complex combinations of features reflecting the interests of diverse parties. These factors, combined with the resulting adherence costs and the infringement of autonomy in the partaking organization, contribute to lowering the level of standards adoption – that is partially mitigated by the return on investment. Moreover, the larger the scope a standard aims to serve, the more complex and difficult its development process can be. In some cases, de facto standards are developed in a spontaneous manner, e.g., because a small group of people have developed a solution that is sound and timely with respect to a certain requirement. Such de facto standards are agreed to in the context of a community of practice, and this makes them less effective when considered in the context of a (global) research data infrastructure dealing with data from many communities of practice to serve the needs of many communities of practice.

Developing comprehensive approaches is complex due to the fact that data interoperability is a problem that goes beyond technical aspects. Data interoperability approaches, to be complete, should reconcile all the differences arising between data providers and data consumers with respect to organizational, semantic, and technical characteristics governing their 'exchange' of data. The majority of existing solutions focus on technical aspects only with very limited efforts and guidelines being oriented to reconcile differences at the organizational level, probably because this domain presents more complexities than others do.

4 RESEARCH DIRECTIONS PROPOSED

There are several fundamental factors guaranteeing that data interoperability is an issue that will continue to occupy stakeholders and practitioners in research data infrastructures and beyond for a long time to come. Among them, the major ones are social factors and complexity factors.

As far as social factors are concerned, data interoperability is a fundamental issue affecting collaboration and data sharing. It involves (i) finding the needed data; (ii) convincing data owners to share data by (a) showing the global and local opportunities and benefits resulting from this action and (b) guaranteeing that their concerns on data sharing (e.g., sharing policies, privacy, provenance, attribution) will be addressed; (iii) convincing data consumers to rely on shared data for real life investigations by guaranteeing authoritative and shared mechanisms for effective and extensive data quality (Batini & Scannapieco, 2006) assertion and communication.

When dealing with complexity factors, even though data interoperability – including the problems it subsumes such as data integration and exchange – is amongst the older research topics in data management, no 'silver bullet' solution yet exists, nor is it expected that such a 'silver bullet' solution will be developed in the near future. However, a rich array of approaches and solutions has been developed in different domains over the last forty years. These approaches and solutions represent a valuable resource that is worth sharing across the boundaries of the domain or discipline that the single approach or solution has been initially developed for so that the status of the practices promoting data interoperability can be further enhanced.

Given the factors guaranteeing that data interoperability is a difficult yet fundamental issue of (global) research data infrastructures, it is fundamental to have a shared and participative strategy to resolve it. This strategy is based on four elements: (i) a shared and comprehensive interoperability framework, (ii) adaptable information objects, (iii) an infrastructure offering interoperability-enabling tools and services whose strengths and weaknesses are a-priori known, and (iv) a 'sandbox' promoting the development, testing, and certification of new interoperability-enabling tools and services.

The interoperability framework is a comprehensive model to be used to characterize the data interoperability problem facets in a systematic way as well as to characterize the existing and forthcoming solutions and approaches. In fact, the lack of a common understanding of what data interoperability is and the absence of a shared language for describing data interoperability and its features are among the most important factors contributing to making this problem a challenging one. This interoperability framework will not be developed from scratch; rather it will be developed by relying on existing ones, e.g., the EIF (European Commission, 2004), with the goal to extend it in order to capture the entire spectrum of aspects characterizing data interoperability in the context of (global) research data infrastructure(s). The framework will cover the whole interoperability problem space from the technical layer to the organizational layer including semantic aspects. The benefits resulting from the development and usage of such a shared model include (a) a unified way to characterize and thus compare problems and solutions, (b) a structure to develop a comprehensive and clear picture of the state-of-the-art with respect to this challenging topic, and (c) a model to drive future effort and initiatives towards enhancing the data interoperability solutions portfolio, e.g., by giving attention to interoperability aspects at the organizational level. The framework might be contextualized (e.g., adapted, specialised) according to the specific needs of a given community of practice; however, it must remain the lingua franca making the developments performed in the context of the community of practice interoperable with the rest.

'Adaptable information objects' are information objects representing research data that are equipped with contextual information enabling the reuse of original data in contexts different from those that the object has initially been created for. Such objects provide different users with diverse 'views' over the same original data. To some extent, such type of object implements the notion of a boundary object (Star & Griesemer, 1989), i.e., a

physical or virtual object that (a) is conceived to live in multiple social worlds by assuming different identities in each world and (b) is conceived to allow coordination without consensus. Adaptable information objects are expected to evolve during their lifetime because of their reuse in multiple domains while guaranteeing attribution, lineage, privacy, and quality with respect to the original data.

The possibility of a data infrastructure offering interoperability-enabling tools and services is a real service that has to be an integral part of any (global) research data infrastructure offering. The larger the array of tools and facilities in place to support data interoperability, the more effective the research data infrastructure will be with respect to data sharing and reuse. However, it is neither feasible nor probable to have a single research data infrastructure that is a priori capable of supporting any data interoperability case. Rather, the envisaged infrastructure will be equipped with a comprehensive set of tools and approaches promoting a 'minimal' level of data management facilities oriented to interoperability and promoting data sharing, discovery, and consumption across the boundaries posed by the intrinsic characteristics of the data and those of the data sources they belong to. These basic facilities are expected to be complemented with specific services and resources, e.g., mappings, tailored to enhance the level of integration between a specific set of data sources and with respect to the interoperability needs of a specific application scenario.

It is fundamental to have a rich and flexible development environment where innovative approaches can be tested – the sandbox. Furthermore, such a sandbox is expected to be used to create new interoperability settings that call for innovative solutions, namely data sources (or samples of them) with certain characteristics with respect to typology, completeness, consistency, precision, and whatever others have to be integrated to some extent. Such a sandbox represents an effective tool where practitioners can both verify the effectiveness of existing approaches and develop completely new and 'certified' approaches – even built by combining existing ones through innovative workflows – aiming at resolving data interoperability issues. Such a certification aspect is particularly important and can be achieved if, and only if, the recommendations above are in place, i.e., the interoperability framework is the model driving interoperability developments including the description of the characteristics of a given solution while the data infrastructure is in place to host and make available such solutions.

5 REFERENCES

Anthes, G. (2010) Happy Birthday, RDBMS! *Communications of the ACM 53*(5), pp 16-17.

Asuncion, C. & van Sinderen, M. (2010) Pragmatic interoperability: A systematic review of published definitions. *Enterprise Architecture, Integration and Interoperability, IFIP Advances in Information and Communication Technology*, Springer, pp 164-175.

Batini, C. & Scannapieco, M. (2006) *Data Quality: Concepts, Methodologies and Techniques*. Springer: Berlin, Heidelberg.

Borgman, C. L. (2010) Research Data: Who will share what, with whom, when, and why? *China-North America Library Conference, Beijing. China*. Retrieved from the World Wide Web, May 14, 2013: http://works.bepress.com/borgman/238.

Borgman, C. L. (2011) The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology.*

Candela, L., Akal, F., Avancini, H., Castelli, D., Fusco, L., Guidetti, V., et al. (2007) DILIGENT: integrating Digital Library and Grid Technologies for a new Earth Observation Research Infrastructure. *International Journal on Digital Libraries 7*(1-2), pp 59-80.

Candela, L., Castelli, D., & Thanos, C. (2010) Making Digital Library Content Interoperable. *Digital Libraries - 6th Italian Research Conference, IRCDL 2010, Padua, Italy*. Revised Selected Papers, pp 13-25.

Candela, L., Castelli, D., Pagano, P., & Simi, M. (2005) From Heterogeneous Information Spaces to Virtual Documents. Digital Libraries: Implementing Strategies and Sharing Experiences. *Proceedings of the 8th International Conference on Asian Digital Libraries, ICADL 2005*, *Bangkok, Thailand*, Springer.

De Roure, D., Goble, C., & Stevens, R. (2009) The design and realisation of the myExperiment Virtual Research Environment for social sharing of workflows. *Future Generation Computer Systems* (25), pp 561-567.

European Commission (2010) A Digital Agenda for Europe - Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. Brussels: European Commission.

European Commission (2004) European Interoperability Framework for Pan-European eGovernment Services. Luxembourg: European Commission.

European Parliament, Council (2007) Directive 2007/2/EC of the European Parliament and of the Council of 14. Establishing an Infrastructure for Spatial Information in the European Community (INSPIRE).

Geraci, A. (1991) *IEEE Standard Computer Dictionary: Compilation of IEEE Standard Computer Glossaries*. IEEE Press.

Halevy, A., Rajaraman, A., & Ordille, J. (2006) Data Integration: The Teenage Years. *Proceedings of the 32nd International Conference on Very Large Data Bases* (VLDB '06), pp 9-16.

Haslhofer, B., & Klas, W. (2010) A Survey of Techniques for Achieving Metadata Interoperability. *ACM Computing Survey 42*(2).

Heath, T. & Bizer, C. (2011) *Linked Data: Evolving the Web into a Global Data Space*. Morgan & Claypool.

Heiler, S. (1995). Semantic interoperability. *ACM Computing Survey 27*, pp 271-273.

Hey, T., Tansley, S., & Tolle, K. (2009) *The Fourth Paradigm - Data-intensive Scientific Discovery*. Microsoft Research.

Kolaitis, P. G. (2005) Schema Mappings, Data Exchange, and Metadata Management. Chen Li (Ed.): *Proceedings of the Twenty-fourth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems,* pp 61-75, Baltimore, Maryland, USA: ACM.

Lagoze, C., & Van de Sompel, H. (2008) Open Archives Initiative Object Reuse and Exchange User Guide - Primer. Retrieved from the World Wide Web, May 14, 2013: http://www.openarchives.org/ore/1.0/primer

Lave, J. & Wenger, E. (1991) *Situated Learning: Legitimate Peripheral Participation*. New York, NY: Cambridge University Press.

Lenzerini, M. (2002) Data integration: a theoretical perspective. *Proceedings of the 21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems* (PODS '02), pp 233-246. New York, NY, USA.

Moreau, L. (2010) The Foundations for Provenance on the Web. *Foundations and Trends in Web Science 2* (2-3), pp 99-241.

Nowakowski, P., Ciepiela, E., Harężlak, D., Kocot, J., Kasztelnik, M., Bartyński, T., et al. (2011)The Collage Authoring Environment. *Procedia Computer Science 4*, pp 608-617.

Paepcke, A., Chang, C. K., Winograd, T., & García-Molina, H. (1998) Interoperability for Digital Libraries Worldwide. *Communications of the ACM 41*, pp 33-42.

Park, J., & Ram, S. (2004) Information Systems Interoperability: What Lies Beneath? *ACM Transactions on Information Systems 22*, pp 595-632.

Rahm, E., & Bernstein, P. A. (2001) A survey of approaches to automatic schema matching. *The VLDB Journal - The International Journal on Very Large Data Bases 10*(4), pp 334-350.

SDMX Initiative (n.d.) SDMX - Statistical Data and Metadata Exchange. Retrieved from the World Wide Web, May 14, 2013: http://sdmx.org

Stapleton, L. K. (2011) Taming big data. *IBM Data Management* (2).

Star, S. & Griesemer, J. (1989) Institutional ecology, 'Translations' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology. *Social Studies of Science 19*(3), pp 387-420.

Van Gorp, P. & Mazanek, S. (2011) SHARE: a web portal for creating and sharing executable research papers. *Procedia Computer Science 4*, pp 589-597.

Wegner, P. (1996) Interoperability. *ACM Computing Survey 28*, pp 285-287.

Wiederhold, G. & Genesereth, M. (1997) The Conceptual Basis for Mediation Services. *IEEE Expert: Intelligent Systems and Their Applications 12*(5), pp 38-47.

(Article history: Available online 1 July 2013)