Data-Centricity and Services Interoperation

Richard Hull

IBM T. J. Watson Research Center, New York, USA
hull@us.ibm.com

Abstract. This position paper highlights three core areas in which persistent data will be crucial to the management of interoperating services, and highlights selected research and challenges in the area. Incorporating the data-centric perspective holds the promise of providing formal foundations for service interoperation that address issues such as providing a syntax-independent meta-model and semantics, and enabling faithful modeling of parallel interactions between multiple parties.

1 Introduction

Services-oriented computing has been evolving, from its roots in orchestration and choreography to the recent tremendous growth in usage stemming from the Software-as-a-Service paradigm and the pragmatism of open REST APIs. Strong notions of "type" have been dropped in favor of message-based API's that refer to data objects with flexible and possibly nested structure. However, as we embrace a world of rich and rapidly created combinations of SaaS-based services from massive numbers of third-party sources, we face challenges of ontology mismatch, entity resolution, and correlation confusion. These challenges must be addressed if we are to find formal and syntax-independent abstract models of service interoperation, systematic design methods for large-scale service compositions, and approaches to support intuitive and formal reasoning about them. Solving these challenges will involve multiple techniques and new advances, but a key element will rely on a shift towards *data-centricity*, that is, enabling data to be at the heart of conceptual modeling, design, and reasoning for interoperating services.

This position paper highlights the need for data-centricity, and overviews research progress and challenges in the area. In particular, the next three sections overview issues and relevant research to date on shared vocabularies and ontologies, entity resolution, and entity correlation; and the concluding section highlights selected research challenges raised and and opportunities enabled by incorporating a data-centric perspective into services interoperation.

2 Shared Vocabulary

Service interoperation involves the exchange of information between services; this is predicated on the assumption that the services involved have an agreed upon meaning for the information. As outlined briefly below, the general solution to enabling rapid

S. Basu et al. (Eds.): ICSOC 2013, LNCS 8274, pp. 1–8, 2013.

compositions of services from multiple sources will require access to ontology mappings, which in turn will rely on both stored information and description-logic style reasoning.

Under the traditional solution, a standard is established that a large body of services are to follow, e.g., see the work of the United Nations Economic Commission for Europe that maintains standards for Electronic Data Interchange (EDI) relating to commercial activities [30]. In the services realm, the SA-WSDL standard [20] enables specification of the ontologies to be used for interpreting values of parameters mentioned in service APIs. Importantly, recent work such as [21] is developing extensions of SA-WSDL to apply to REST APIs, and enabling matchmaking techniques to be applied on them.

In many application areas, however, it is not possible to enforce a universal standard with regards to the vocabulary or ontology used by services. This is acknowledged in the data management literature, for example, [9] provides a survey of problems and techniques for integrating data stored according to different vocabularies and ontologies. Citation [16] argues that we must live with such heterogeneity in the healthcare domain, and [18] argues similarly for education. More generally, [10] provides a comprehensive discussion of techniques to semi-automatically develop mappings between ontologies, and how the results are used in a variety of applications.

From the perspective of service interoperation, simply having the ability to semi-automatically compute ontology matchings is not sufficient. In particular, mechanisms are needed to access such matchings, either from a locally stored ontology or through a service. A rich example of the latter is found in OntoCAT [2], which provides APIs for going between multiple ontologies in the bioinformatics field. Tools such as the Karlsruhe Ontology (KAON) infrastructure [17] and the Ontology Mapping Store [25] provide generic access to ontology mappings.

To summarize, although ontology mappings are not in practical use to support modern service interoperation, much of the foundational research and several research tools have been developed in recent years. In the coming years requirements from industry will help to determine the application, business models, and evolution of these techniques.

3 Entity Synonym Repositories

Entity resolution, that is, the problem of extracting, matching, and resolving occurrences of entity names in structured and unstructured data has a history going back to the 1950's. This topic has become important again in recent years because of the interest in so-called "big data" and applications in advertising, marketing, and personalized services that attempt to mine social data for useful, entity-specific information. A survey of the field, including recent advances that use advanced machine learning techniques, is provided in the recent tutorial [11].

On the positive side, many successful techniques are in place to achieve relatively accurate entity resolution. But in practical systems that need a very high degree of precision it is typical to augment the automated techniques with manual validation activity. Although not scalable in a true sense, this can provide a pragmatic approach to incrementally build up near-certain information that augments the automated techniques.

To illustrate, consider applications that attempt to find sales leads for business-to-business (B2B) companies, by searching through news articles, blogs, and other social media for events that suggest that a given company might be helped through the purchase of a given product. Although there are pseudo-standards for company names, e.g., the Dunn and Bradstreet database, companies are often referred to by a handful of synonyms in the media. As a result, a viable service in this space will augment automated techniques by storing a dictionary of manually determined synonyms. Whether this is stored as part of the service, or is accessed from an external service, it is nevertheless a persistent data store.

Increasingly, interoperating services will be accessing unstructured data and/or data from multiple repositories. As a result they will need to rely on entity resolution, and on associated repositories of entity synonyms. It is likely that multiple proprietary and externally accessible services will become available that provide access to synonym repositories, to enable uniform entity reference across interoperating services.

4 Managing Entities across Services

In many service interoperation scenarios there are multiple entities, either physical or conceptual, that are being managed or manipulated through time by different services. This requires precise management of the relationships between the entities, called *correlation* in the early literature on orchestration and choreography of web services. That early literature does not provide mechanisms for explicitly modeling or specifying such correlations. Recently, [29] has developed a framework for such explicit modeling, based on the notion of business artifacts This includes a declarative language for intuitive yet systematic specification of correlations across entities along with constraints on the correlations, and also supports the possibility of formal reasoning about them. This section briefly explores some of the ways that the business artifact perspective can support the management of correlations.

Since their introduction in 2003 [26,19], business artifacts have been shown useful for several aspects of business process modeling. Briefly, a business artifact type is used to represent a class of (physical or conceptual) entities that evolve as a business process unfolds. An artifact type includes both an *information model*, that provides room to hold (possibly nested) data about an artifact instance that may change over time, and a *lifecycle model*, that holds a specification of the possible ways that the artifact instance might evolve. (The lifecycle model might be specified in a procedural paradigm such as finite state machines or Petri nets, or a more declarative paradigm such as DECLARE [27] or Guard-Stage-Milestone [13,8]). Applications of the artifact perspective include enabling a cross-silo view of business processes [28], providing a coherent integrated view of multiple similar but different business processes [7], business process performance monitoring [23] and enabling data-centric service interoperation [15]. Further, artifacts are closely related to "cases" in the sense of Case Management, and the recent OMG Case Management Modeling and Notation (CMMN) standard [6] embodies a merging of these two streams of conceptual modeling [24]. We believe that the artifact approach will bring numerous advantages to our understanding, design, and deployment of service compositions, including those just mentioned. We focus here on the

issue of correlation because it provides the single most easily motivated application of the artifact perspective to services interoperation.

To illustrate the basic form of correlation, consider the "make-to-order" example as described in [14] and elsewhere. In this scenario, a "Customer Purchase Order" is sent to an "assembler", who in turn creates several "Line Items" that must be obtained in order to fulfill the customer order. The assembler researches the problem of which suppliers to use to obtain the line items, and eventually creates multiple "Supplier Purchase Orders", each of which may request multiple line items from a given supplier. Furthermore, in some cases a supplier may reject a Supplier Purchase Order, which means that the affected Line Items must be grouped again to generate additional Supplier Purchase Orders. Importantly, the specific relationships evolve over time, e.g., as, Line Items are created, as Supplier Purchase Orders are created, and as Supplier Purchase Orders are fulfilled or rejected. It is easy to see that in the general case, there are *1:n* relationships between Customer Purchase Orders and Line Items. If we consider Line Items and Supplier Purchase Orders over time, then *n:m* relationships may arise, e.g., if a Supplier Purchase Order is rejected and then one of its Line Items is placed into a second Supplier Purchase Order. Business artifacts were used in 2005 in support of an application with similar kinds of correlations involving 1000's of interrelated objects [5]. The application, based on the pharmaceutical drug discovery process, was centralized in that study, but would most likely be distributed across multiple services if developed today.

As mentioned above, [29] shows that a natural approach for explicitly modeling correlations in a services interoperation context is through the use of business artifacts. In the make-to-order example, three artifact types can be used, one each for Customer Purchase Order, Line Item, and Supplier Purchase Order. It is straightforward in this context to maintain the correlations between artifact instances, e.g., in a deployment where each artifact type is maintained by a separate service. This approach can be extended to situations where one or more artifact types is managed by multiple services, rather than by just one. The approach can be used in a choreography-based setting, as illustrated in [29], or in an orchstration-based setting such as [15,22].

A richer form of correlation arises when entities can be split apart or merged. One example of this arises in the context of Collaborative Decision Processes. These are processes that involve multiple stakeholders who together explore a variety of ideas and initiatives in order to reach a (typically multi-faceted) decision over a period of time (e.g., weeks or months). An example application is when members of a community decide on the characteristics that should be embodied in a new shopping mall. This may include several investigations into traffic impact, watershed impact, etc., and also the exploration of numerous alternatives. In some cases these initiatives may split (e.g., consideration of a recreational shop for the mall may split into considerations for adult recreation and for youth recreation) or iniatives may merge (e.g., separate initiatives around a movie theater and something for art lovers might merge into an initiative for an independent film theater). As described in [31], it is natural to model and implement such processes using the business artifact approach, using an artifact instance for each idea or initiative that is explored. This simplifies the use of multiple interoperating services when implementing the core of the decision process, and also when incorporating new services, e.g., to solicit opinions from a crowd or to conduct polls.

An application area that involves a broad array of stakeholder enterprises and where conceptual entities can transform is in tracking food in the supply chain "from farm to fork". For example, with a potato in a frozen stew, multiple interrelated "lots" (i.e., collections of goods treated as a unit for shipping or processing activities) come into being, as the result of combining potatoes from different harvests, mixing different ingredients, and finally the packaging and delivery. Monitoring the overall process and enabling adjustments to it (including recalls) is vastly simplified if a data model is deployed as the backbone for the interoperation. In this case, the lots are naturally modeled as business artifacts, and systematic correlation between lots is easily managed, even as the underlying goods are processed.

As illustrated by the examples above, the artifact perspective can provide a natural and comprehensive modeling framework for managing entities and their correlations in service compositions. At the core of the approach is the understanding that the various services are manipulating a common set of conceptual entities. Multiple approaches can be used for the actual storage of associated information, e.g., maintaining full artifact instances within a single service, distributing their storage across multiple services, or passing partially completed artifact instances from one service to another (cf. [1]).

5 The Challenges Ahead

Data has always been fundamental in service interoperation, but until recently it has not been emphasized in the research on conceptual models, in specification languages, or in the study of foundations. With the dramatic increase in SaaS offerings and the anticipation that many business processes will be performed using rapidly created service compositions, the systematic and intuitively natural management of the data aspect along with the process aspect will become essential. As discussed above, there are three main components to the data aspect, namely, access to ontology matchings, access to synonym repositories, and the management of entity correlations. Three broad research areas are now highlighted.

One broad challenge area concerns extensions of "keep it simple" approaches such as REST and noSQL (e.g., as embodied by JSON) to incorporate data more explicitly. As noted above, the piece-parts to support such extensions are now available in the research literature. But finding ways to combine them in simple ways that become widely adopted remains open. Here the ontology matching and entity name synonyms can be viewed as more-or-less stand-alone services giving access to essentially static data, In contrast, entity correlation involves dynamic data and traditional database considerations such as transactional consistency, maintenance of equivalence across copies of data held in different services, and preservation of integrity constraints across updates. Such issues become more intricate in cases where the underlying implementation for one or more services is parallelized and/or distributed for performance reasons.

A second challenge area involves developing mechanisms for reasoning about service interoperation, reasoning that incorporates the data aspect along with the process aspect. Promising formal work in this direction is provided by, e.g., [4], which develops formal verification techniques for distributed artifact-based systems, and by, e.g., [12] and related papers, which develop a rich theoretical basis for formal verification

of systems that involve ontology, data, and the evolution of data reminiscent to that found in artifact systems. The latter work is especially relevant to service interoperation where ontology matching is involved, since those matchings may include a combination of fixed data and description logic based reasoning. More broadly, this work will help to pave the way for practical (possibly semi-automatic) verification tools that enable faster and more automated development and testing of service compositions, and also for debugging systems that can help to explain why errors are occurring.

A third area where the data-centric perspective may have useful application for service interoperation is to address concerns around foundations for the orchestration and choreography standards, as raised in [3] and elsewhere. In particular, the artifact perspective holds the promise of providing a formal grounding for both orchestration and choreography, precisely because it can faithfully represent not only the process aspects but also the data aspects. Reasoning about the interactions between entities can be explicit, as already demonstrated in [29]. More broadly it appears that these techniques can be extended to reason about both the parties involved in a service composition and the conceptual entities being manipulated on their behalf. The artifact perspective can be extended to provide an explicit formal meta-model underlying the BPEL and WS-CDL standards, which would enable rich styles of formal verification for specifications in those standards, styles that incorporate the data aspects as well sa the process aspects. Multi-party interactions, and in particular managing multiple interactions that may happen in parallel, can be formalized and studied, as illustrated by the work on entity correlation [29] described above, and by research on artifact-centric interoperation hubs [15]. Finally, the artifact perspective may provide a workable basis for developing a theory around the integration of multiple service compositions, including an understanding of compositions of compositions, an area that will be increasingly important as the use of service compositions continues to grow.

References

1. Abiteboul, S., Benjelloun, O., Milo, T.: The Active XML project: An overview. Very Large Databases Journal 17(5), 1019–1040 (2008)
2. Adamusiak, et al.: OntoCAT simple ontology search and integration in Java, R and REST/-JavaScript. BMC Bioinformatics 12(218) (2011), http://www.biomedcentral.com/1471-2105/12/218
3. Barros, A., Dumas, M., Oaks, P.: Standards for web service choreography and orchestration: Status and perspectives. In: Bussler, C.J., Haller, A. (eds.) BPM 2005. LNCS, vol. 3812, pp. 61–74. Springer, Heidelberg (2006)
4. Belardinelli, F., Lomuscio, A., Patrizi, F.: Verification of agent-based artifact systems. CoRR, abs/1301.2678 (2013)
5. Bhattacharya, K., et al.: A Model-driven Approach to Industrializing Discovery Processes in Pharmaceutical Research. IBM Systems Journal 44(1) (2005)
6. BizAgi, Cordys, IBM, Oracle, SAP AG, Singularity (OMG Submitters), Agile Enterprise Design, Stiftelsen SINTEF, TIBCO, Trisotech (Co-Authors).: Case Management Model and Notation (CMMN), FTF Beta 1 (January 2013), OMG Document Number dtc/2013-01-01, Object Management Group

7. Chao, T., et al.: Artifact-based transformation of IBM Global Financing. In: Dayal, U., Eder, J., Koehler, J., Reijers, H.A. (eds.) BPM 2009. LNCS, vol. 5701, pp. 261–277. Springer, Heidelberg (2009)
8. Damaggio, E., Hull, R., Vaculín, R.: On the equivalence of incremental and fixpoint semantics for business artifacts with guard-stage-milestone lifecycles
9. Doan, A., Halevy, A.Y.: Semantic-Integration Research in the Database Community. AI Magazine 26(1) (2005)
10. Euzenat, J., Shvaiko, P.: Ontology matching. Springer, Heidelberg (2007)
11. Getoor, L., Machanavajjhala, A.: Entity resolution: Theory, practice & open challenges. Proc. of the VLDB Endowment (PVLDB) 5(12), 2018–2019 (2012)
12. Hariri, B.B., Calvanese, D., De Giacomo, G., Deutsch, A., Montali, M.: Verification of relational data-centric dynamic systems with external services. In: Proc. Intl. Conf. on Principles of Database Systems (PODS), pp. 163–174 (2013)
13. Hull, R., et al.: Introducing the guard-stage-milestone approach for specifying business entity lifecycles. In: Bravetti, M., Bultan, T. (eds.) WS-FM 2010. LNCS, vol. 6551, pp. 1–24. Springer, Heidelberg (2011)
14. Hull, R., et al.: Business artifacts with guard-stage-milestone lifecycles: Managing artifact interactions with conditions and events. In: ACM Intl. Conf. on Distributed Event-based Systems, DEBS (2011)
15. Hull, R., Narendra, N.C., Nigam, A.: Facilitating workflow interoperation using artifact-centric hubs. In: Baresi, L., Chi, C.-H., Suzuki, J. (eds.) ICSOC-ServiceWave 2009. LNCS, vol. 5900, pp. 1–18. Springer, Heidelberg (2009)
16. Iroju, O., Soriyan, A., Gambo, I.: Ontology matching: An ultimate solution for semantic interoperability in healthcare. International Journal of Computer Applications 51(21), 7–14 (2012)
17. KAON2 Home Page (2005), http://kaon2.semanticweb.org/
18. Kiu, C.-C., Lee, C.-S.: Ontology mapping and merging through ontodna for learning object reusability. Educational Technology & Society 9(3), 27–42 (2006)
19. Kumaran, S., Nandi, P., Heath III, F.F. (T.), Bhaskaran, K., Das, R.: Adoc-oriented programming. In: SAINT, pp. 334–343 (2003)
20. Farrell, L., Lausen, H.: Semantic Annotations for WSDL and XML Schemas. W3C Recommendation (August 2007), http://www.w3.org/TR/sawsdl/
21. Lampe, U., Schulte, S., Siebenhaar, M., Schuller, D., Steinmetz, R.: Adaptive matchmaking for restful services based on hrests and microwsmo. In: Proceedings of the 5th International Workshop on Enhanced Web Service Technologies, WEWST 2010, pp. 10–17. ACM, New York (2010)
22. Limonad, L., Boaz, D., Hull, R., Vaculín, R., Heath, F.(T.): A generic business artifacts based authorization framework for cross-enterprise collaboration. In: SRII Global Conference, pp. 70–79 (2012)
23. Liu, R., Vaculín, R., Shan, Z., Nigam, A., Wu, F.: Business artifact-centric modeling for real-time performance monitoring. In: Rinderle-Ma, S., Toumani, F., Wolf, K. (eds.) BPM 2011. LNCS, vol. 6896, pp. 265–280. Springer, Heidelberg (2011)
24. Marin, M., Hull, R., Vaculín, R.: Data centric BPM and the emerging case management standard: A short survey. In: La Rosa, M., Soffer, P. (eds.) BPM 2012 Workshops. LNBIP, vol. 132, pp. 24–30. Springer, Heidelberg (2013)
25. Marte, A., Fuchs, C.H.: OMS - Ontology Mapping Store (2013), Available on Source Forge, http://sourceforge.net/projects/om-store/

26. Nigam, A., Caswell, N.S.: Business artifacts: An approach to operational specification. IBM Systems Journal 42(3), 428–445 (2003)
27. Pesic, M., Schonenberg, H., van der Aalst, W.M.P.: Declare: Full support for loosely-structured processes. In: IEEE Intl. Enterprise Distributed Object Computing Conference (EDOC), pp. 287–300 (2007)
28. Strosnider, J.K., Nandi, P., Kumaran, S., Ghosh, S., Arsanjani, A.: Model-driven Synthesis of SOA Solutions. IBM Systems Journal 47(3) (2008)
29. Sun, Y., Xu, W., Su, J.: Declarative choreographies for artifacts. In: Liu, C., Ludwig, H., Toumani, F., Yu, Q. (eds.) ICSOC 2012. LNCS, vol. 7636, pp. 420–434. Springer, Heidelberg (2012)
30. United Nations Economic Commission for Europe (UNECE). Introducing UN/EDIFACT (2013), `http://www.unece.org/trade/untdid/welcome.html`
31. Vaculín, R., Hull, R., Vukovic, M., Heath, T., Mills, N., Sun, Y.: Supporting collaborative decision processes. In: Intl. Conf. on Service Computing, SCC (2013)