1 Information Integration in Context[1]

I'd like to circulate this among the panel members fairly soon. In particular we should ask for citations Ken, Jim, Robert – should we say more about the perceived state of the art?

Let us make the assumption – and it is a big assumption – that the tools for mapping and matching have been perfected and that all that remains is to use these excellent tools. To what extent would they tools fulfil the needs of the users of integration systems?

For the purposes of this discussion we shall divide the users of these tools into two classes: one is the *brokers*, the people who use the tools to generate integrated data sets, the other is the *consumers*, the end users of these data sets. They are not always distinct: there are often cases one person doing "one-off" integration for the sole purpose of answering a single question.

1.1 Training

Tim, Jim, Deborah, Laura, Victor?

No-one believes that there is a silver bullet for information integration. We shall always need brokers. But what kind of skills should they have? We expect that data integration tools will become sufficiently simple that they will not need a PhD in Computer Science to use them. However, no-one thinks that they can be used without some kind of understanding of data models and how integration tools operate on them. More importantly we expect the brokers to be domain experts. So will these people, for example, be biologists with some knowledge of CS will they be computer scientists who have learnt some biology or will we require both? Our current belief is that we need both, moreover special training should be introduced into the relevant curricula in data integration, and that understanding the semantics of data models/schemas/ontologies should be an important part of this training.

1.2 Forms of integration

Cite astrodas

The current view of integration is to combine n sources into one, but within this general prescription there is a wide variety of forms of integration. If we assume (usually incorrectly) that the sources are static there is the issue of whether we are trying to produce a comprehensive integration of all the sources, as is done in some scientific databases or are we trying to build something that will answer a narrow, focused set of queries. Are the demands on the brokers that they produce something as quickly as possible (best effort) or do they have the time to do a thorough job. Another issue that has not been properly addressed in integration tools is that of time. In demographic and health data integration of several data sets from different periods in time is often used to get a "longitudinal" study of how data such as population or the incidence of a disease changes over time. Synchrony is also important in ordinary integration. There are many stories of disasters that have resulted from integration with stale data.

1.3 Security and privacy

Need to co-ordinate with Jag's session Barbara?

Does integration compromise security? This goes both ways. Sometimes integration is used to ensure privacy. For example, on person identifiers and then erasing those identifiers and perhaps some other identifying data to obtain aggregate data for medical research is common practice. Conversely trying to combine anonymized data is used – sometimes benignly – to identify people with medical conditions. Security can also be an issue in blocking explanations and provenance (see below.)

[1]Contributors: Barbara Blaustein, Peter Buneman, Sarah Cohen-Boulakia, Robert Chadduck, Yi Chen, Laura Haas, Tim Finin, James French, Yannis Ioannidis, Subbarao Kambhampati, Deborah McGuinness, Victor Markowitz, Wang-Chiew Tan, Kenneth Thibodeau and Gary Walter

1.4 Source selection and discovery

Finding and assessing the relevant sources is not always straightforward, and once the sources have been identified, it is often the case that several sources will provide the "same" data so that the broker is faced with the problem of source selection. Here is a list of some of the factors that may govern source selection:

Expanded from Subbarao's comments and Yanni's list

- authorship – who generated the data in the case that it was a person?
- provenance – where did the data come from, in the case that it was copied from somewhere?
- completeness – does the source provide comprehensive coverage of the data of interest?
- performance – how efficiently can the tools interact with the source
- format – integration tools are, of course designed to deal with different formats, but there are always issues of how easily the format can be understood (for example, is further parsing of text fields or conversion of numeric fields needed.)
- cost
- cleanliness – does the data conform to constraints that may be assumed by the schema used in the integration rules?
- freshness – how up-to-date is the source?

Many of these topics fall under the general heading of *data quality* or trust – a topic that requires further investment of research effort.

1.5 Source understanding

How do brokers understand other people's data? This is always a problem in data integration. The overarching requirement is that the sources should be documented at all levels: the description of the units used in some numeric field is as important as the description of the schema; moreover (see below) it is important that this information is carried through integration.

Several people have noted that looking at the data is often a better entry into understanding a data source than looking at the schema. In fact some mapping tools have a facility for doing this.

Laura - a citation?

1.6 Databases or Ontologies?

Tim, Deborah?

Ontologies have recently emerged as an alternative to databases for certain kinds of knowledge representation. Ontologies may permit a more flexible approach to designing data sets – especially in their initial evolution – while databases still provide the basic technology needed for efficient access to large shared data sets. Is one more appropriate than the other for integration? There is a close relationship between database schemas and ontologies, so the likelihood is that integration tools developed for one of these will be transferable to the others. Nearly all the issues discussed in this section on context are applicable to both.

1.7 Data consumers

Most interfaces for consumers are engineered – as they should be – to present the data in a form that is most understandable. Typical examples include the sequence displays one finds in genetic databases and the pictorial displays of astronomical data . Unfortunately the data display is far too often the *only* end product of the integration. There are two other important extensions that are needed.

citations

First the product of any integration service is likely, if it is of sufficient quality, to be needed as input to some further analysis tool or some other integration service. Therefore it is also essential that the data is made available in a some clean understandable format with full documentation. Many integration tools do this, but given the further demands we are about to place upon data integration tools, this is an entirely non-trivial task.

This was the point made by Dave Maier

The second extension is that of *explanation*. Consumers need more than a well engineered presentation of the data. They need the ability to "drill down" or expand parts of the data in order to understand why it is there. This is a sufficiently important addition to current data integration tools that it requires special attention.

Tim's point about data being seen at various levels of granulaity.

1.8 Explanation

.

Laura, Jim??

There are cases in which the results of some integration task were dismissed by the consumers, simply because they did not know how the integration had been performed. Worse, the same users then attempted to re-integrate the relevant sources in order to be sure that they knew what the integrated data meant. Some form of explanation of *how* the integration was performed and the relevant contextual information is essential. Two components of explanation are particularly important.

1.8.1 Provenance

In the context of information integration, provenance is a record of how the integrated data was derived. Two general approaches to this general topic have emerged. The first is *workflow* provenance in which one keeps a complete record of how the resulting data set evolved. Workflow provenance involves recording not only the data sets and the transformations that were performed, but also the details of external computations that may have been used. For example Blast searches are frequently used in biological data integration. The second *fine-grain* approach focuses on individual data elements in (or small components of) the integrated data set and asks for the provenance of that element. It may be that providing the provenance of small pieces is simpler than recording an entire workflow. This is certainly true in the case that the integration tools simply caused copying of the data.

Victor – citation?

Wang-chi's, Deborah?

1.9 Documentation, annotation and context

Important documentation is often lost in integration. An integration tool may do a perfectly good job of producing accurate data, but may lose important information, such as dimensional information that is only available in a comment in the source schema. It is essential that such documentation is carried through, as some form of annotation into the integrated result. This is again a non-trivial trivial task. Even determining what documentation in the source schema is relevant to the extracted data is non-trivial and usually requires human judgement. Finally there are external factors. Knowing the purpose for which a data set was constructed or why an integration task was performed is essential in assessing its use for other purposes.

Yi, Wang-Chiew, Victor?

1.10 Compositionality

To return to our earlier point concerning the requirement that data integration tools should produce, in addition to good user interfaces, well-organized data sets that can be used by other systems, perhaps by other integration tools. The key idea here is compositionality. Not only should integration be composable, but so should the systems that convey documentation, context and provenance. This is one of the key challenges in bringing integration technology to market.

1.11 Curated Databases

A substantial number of databases are produced not by integration tools but by humans who constantly use their judgement in the selection and classification of data. This is true of many of the 800 or more molecular biology databases and it is also true of many on-line reference manuals, gazetteers, business compendia, etc. While automated integration is less relevant to the "curators" of these databases, nearly all the issues that we have discussed in this section on context are equally important to curated databases, and they are equally challenging.

cite Nuc. Acid. Res.