Information Integration of Drug Discovery and Clinical Studies to Support Complex Queries Using an Information Supply Chain Framework

Thomas Kandl
F. Hoffmann-La Roche AG, 4070 Basel, Switzerland
thomas.kandl@roche.com

Nawaz Khan
School of Science and Technology, Middlesex University Hendon, London NW4 BT, UK
n.x.khan@mdx.ac.uk

***Abstract*—Query support of frameworks which are capable of linking and integrating trusted information in disciplines such as clinical studies face many challenges, as usually data are neither semantically consistent nor in a form which can be easily extracted and shared.. Drawing conclusions from clinical studies usually requires integrating data from various sources at various time points and inferring facts from an integrated information set. However, relevant information may not be present at all or not in sufficient detail or quality to support a particular query. In this paper, we present an approach for query support in integrating diverse information sources and how trust can be strengthened in using subject oriented specific Information Supply Chains (ISC). We assess feasibility of query support in an ISC framework using an underlying Information Dependency Relation Model for relevant information sources and how such a framework may facilitate collaboration in supplying information at the right time and with required quality.**

***Index Terms*— Ontology, Information Supply Chain, Clinical Decision Support, Information Integration**

I. Introduction

Answering questions in clinical studies requires integration of information which in many cases may be held in different institutions, in different formats and in various semantic contexts. Such diversity hinders constructing generic ontology that avoids semantic conflicts [1], yet are necessary to enable sharing of data and information across domains, within a context [2,3].

Conclusions about suitability in researching new drugs, even preliminary, require consideration and integration of various trusted information [4,5]. Lineage information provides for trust by describing the relationships between raw data and its transformations after processing has occurred [6]. Details along the processing of original data are equivalently important to understand derived conclusions as their actual values. Hence reuse of scientific information is dependent on how it is trusted and storing provenance information needs to be considered in a suitable framework for collaboration.

Indeed, several design principles for such a framework were described of which capturing associations between sharable resources are considered as key enablers [7]. Yet, proper annotation is inherently a tedious task, especially when such annotation cannot be performed at the time new data is generated. Publishing at source has been hence promoted for trusted information flows [5]. Such a framework captures contextual information at the origin and facilitates proper annotation of data. A similar approach was taken in [8] where the feasibility of so called semantic tagging was demonstrated, in which information was allowed to be added in a free form and then labeling it with appropriate semantic tags that are inferred from present task in the workflow.

Reuse of information is related to its validity and timing – the more that is known about how particular information was derived the more it is trusted [9]. In this respect the concept of just-in-time (JIT) information management is appealing with the aim of providing the right information for the right users at the right time [10,11]. In relation to this, critical success factors in establishing JIT architecture are discussed in [12].

In this study we test feasibility of ISC and have extended the model to apply JIT- strategy in tackling information supply and demand for clinical study. We modeled JIT by the use of information dependency relation that is analogues to Bill of Material (BOM) [13] and have incorporated this structure in personalized views to infer information relationships within the context of clinical study workflows. We have used determination of potential immunogenicity of therapeutical proteins as a case study in particular due to its dependency on timely information integration of different scientific research workflows.

We show how such ISC framework is able to support complex queries by tackling information integration. Specifically we demonstrate the use of personalized Information Dependency Relation (IDR) models and how

doi:10.4304/jsw.9.5.1348-1356

such models are capable of supporting complex queries and trust in obtained results.

The following sections explore shortly information flows and how semantic boundaries hinder integration, followed by a discussion about protocols to support establishing ontology. We then focus on ISCs in more detail and specifically how Information Dependency Relation (IDR) models support rich information integration and allow balancing information demand and supply. Query support of the ISC framework is highlighted and its result discussed in the light of enhancing trust in obtained conclusions. Finally the paper suggests further work of the ISC framework and discusses the contributions of this prototype in complex information integration queries.

II. Trusted Information Integration in Information Supply Chains: Concepts and Measures

In order to derive a conceptual model of trusted information integration in Information Supply Chains (ISC), in the sections below we shortly introduce requirements for querying semantic consistent information, as well as the concept of information completeness to measure the performance of an ISC. We will then demonstrate how these concepts contribute to rich query support

A. Information Flows and Semantic Boundaries

Information flows of two exemplary cases in drug discovery disciplines are shown in Fig. 1. In both cases, information traversed many semantic boundaries (depicted in numbered rectangles) and hence was subject to interpretation in the specific domain.

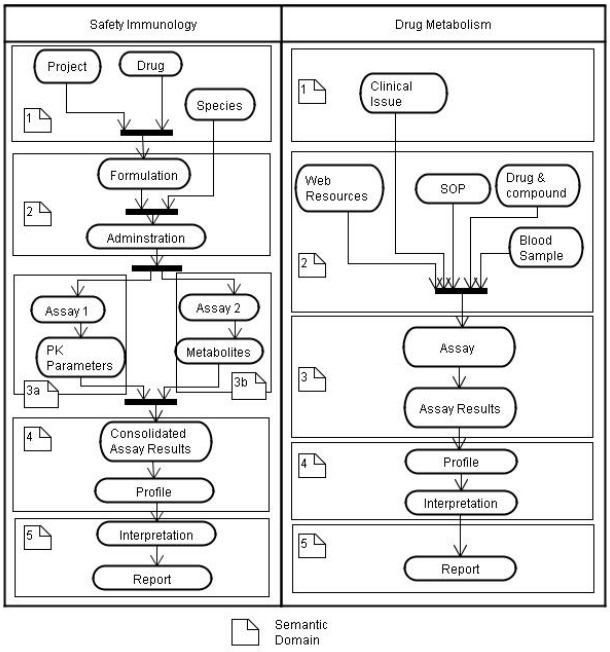

Figure 1. Information flows of Safety Immunology and Drug Metabolism as exemplary cases.

In order to select appropriate assays and apply relevant parameters, laboratories needed specific information about the study or investigation. Such information however was often either not available or not provided at appropriate levels of detail to prepare next steps, leading to assumptions which resulted –in many instances - in biased conclusions.

Also, these labs required many information sources whose relevance and validity played an important role. Relevance was important as information was passed from one domain to the other and was considered valid in certain domains only. Often times it was found that the source information did not provide enough details to verify its validity and hence was either not used or taken as such albeit taking the risk that information had only limited validity. In fact information flows manifested in quite complex networks in which demand and supply were unbalanced as most sources did not provide mechanisms to determine its relevance and validity. Collaboration was hence viewed especially important where information traversed semantic boundaries; however, in general, because of lacking information about interpretations and obtained results, final reports were not trusted, even though they may have come from internal sources, i.e. findings of safety relevant issues in investigational reports performed in internal laboratories, yet not providing sufficiently detailed information about experiment design or data analysis

As often times experimental setup and underlying assumptions were not explicitly stated, and rather external sources referenced, the role of protocols with regards to information flow and how referenced protocols were actually followed was further examined and is discussed in the next section.

B. Role of Protocols

Protocols played a fundamental role in the exemplary cases in that they described the conduct of an experiment and how obtained data were to be processed to conclude a profile. These data, referred to as laboratory records that constituted information, were the result of following a protocol. Validity of a laboratory record was hence dependent on the quality of the protocol description as well as how accurate the protocol was followed (Fig. 2). Experimental design and methods used for interpreting data was complex in both case studies. Hence access to provenance information about laboratory records, how they were obtained, and information about the protocol and also how the step was performed in reality was considered in both cases as highly important. Yet design of a protocol was dependent on relevant information from investigations or studies; those investigations or studies usually had an underlying hypothesis, and hence incomplete or information from a domain different from the applicability of the protocol may have had consequences in concluding obtained data. Hence the quality of the final report reflected information obtained during individual steps of the investigation or study.

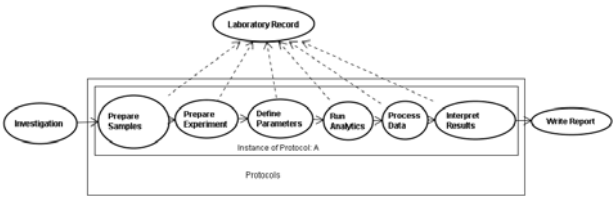

Figure 2. Protocols govern laboratory records which are composed of captured data along a scientific workflow.

Application of protocols and their interpretation is highly specific to a given context and in most cases cannot be applied for another context. To describe protocol contexts within an ISC depends on the successful implementation of ontology and data semantics. Next section explores some of existing technologies that enable ISC to describe the context of protocols.

C. Information Management in ISC

Similar to a supply chain, an ISC addresses information demands in its information network. Development of an ISC framework needs to analyse information sharing requirements, both in terms of issues and opportunities, and failure to do so have similar consequences as SCM: unbalanced demand leads to either information overflow or deficiency. Hence information management of SCM and ISC share the same goal, namely balancing demand and supply. However, information flow and management in SCM differs from ISC in that the former deals with flow of materials whereas the latter deals with flow of information. Nevertheless concepts, methods, goals and philosophy of SCM can greatly improve information sharing and can be – in many cases - directly applied towards ISC [13].

Another difference is that material in SCM needs to be replenished and its inventories controlled whereas information in ISC can actually be copied. Provenance information of information (i.e. how is this information derived), relevance (i.e. is the information still current) and applicability (i.e. how is this information related to a certain problem) governs an ISC.

Drug discovery with its objective of discovering new drugs for the treatment of diseases is heavily dependent on timely and trusted information having access to provenance information [14]. Such information is derived from other information and in many ways resembles a Bill of Materials (BOM). Yet, in order to model ISCs information dependencies need to be understood [15]. Semantic conflicts as outlined in the exemplary cases (Fig. 1) mainly arose because different concepts were applied which could not be easily integrated and mapped to each other. Using agreed ontology to support a common agreed BOM structure, referred to as information dependencies, is believed to lower semantic conflicts. The next section discusses such ontology concepts in the context of ISC framework and how these concepts enable semantic consistency.

D. Information Dependency Relation Models

An ISC addresses information demands in its information network and makes contained information actionable. Access to specific information is required only at specific time points or events along a particular workflow [13]. For example, in order to design a clinical study, access to drug and target information is essential. Relevant patient information is only required when conducting the study. Hence, having access to relevant information at the appropriate time is identified as a key requirement for building information dependencies. Such information dependency relations will of course vary between experts, depending on experience and knowledge of the specific contexts. However, such models can be easily shared and discussed; particular integration of relevant information becomes more trustworthy.

E. Trusted information integration

As outlined above an agreed IDR model helps in building trust in integrating information sources by describing how information is dependent on other information and thereby allowing making complex integration paths "tangible". The level of trust in integration information will depend on supplied information and how relevant such information is for supporting a particular query in which information may be integrated from several information sources. Especially when integrating different sources of information the level of trust will decrease if only limited provenance information is available [4].

Suppose for example an IDR model of a Clinical Investigation, depicted in Figure 3:

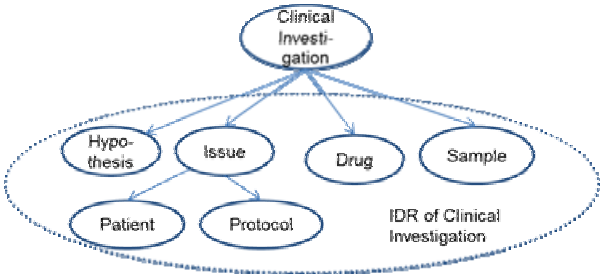

Figure 3. Example of an IDR Model of a Clinical Investigation used in the clinical case study.

This model shows for example, how Protocol relevant information of Clinical Investigations needs to be integrated, namely requiring at least information about the Protocol itself and the Issue to which it is directly connected. If information is available it will satisfy the immediate need for information integration on the level of the Issue, however more information is required in building trust as Clinical Investigations also require information about the Drug, underlying Hypothesis, Samples and information about Patients who exhibit a certain clinical Issue. Only when all information is available, trust in obtained results can be enhanced as single information can be traced to other information it depends on. Hence the level of trust will depend on how

complete available information is to satisfy the information dependencies in an IDR model.

F. Information Completeness and Information Balance

To apply an IDR model to an instance of specific information, it is necessary to measure information completeness (IC). It can be measured as the ratio of all required information instances for that particular information versus actually present. For example, to conclude an investigation, information about Sample, Drug, Clinical Issue and Hypothesis must be present. If all required information is indeed linked to the Investigation then said information is 100% complete. On the other hand, if, for example, Hypothesis is missing the resulting IC is lowered to 75%. The IC of an information instance can be computed as follows:

$$IC = \frac{\sum Ipresent}{\sum Irequired} \tag{1}$$

where Ipresent are information instances that are present and Irequired represent information instances that are required by the underlying IDR model. From the example in Figure, Sample, Drug, Clinical Issue and Hypothesis represent all required information for a Clinical Investigation. The IC of such a Clinical Investigation will therefore approach a value of 1 if it is said to be complete with regards to all its required information.

Since an ISC is made up of information units with their related information dependencies, an IC for an ISC (ICisc) can be computed as the average of all IC values from these information unit instances that are published into the ISC. For example, if an ISC comprises a Clinical Investigation and an Analysis, than the ICisc will be the average of the IC values from all information required by the Clinical Investigation and the ones from the Analysis. Hence, ICisc can be computed as follows:

$$ICisc = \frac{\sum ICn}{Count(IC)} \tag{2}$$

where ICn is the IC value of information unit n in an ISC, and Count(IC) is the total number of all published information into the ISC.

The values for ICisc will range from 0 – no information dependency satisfied – to 1, which signals a complete information set. Such an IC value gives an indication of how complete particular information set of an ISC is and is suitable especially when such information is integrated and used in complex queries. We have used such IC values in determining the demand and supply of required information which is demonstrated in the following section.

G. Related Work

Current literature on information integration considers the issues of information quality, information completeness and information consistency. For example, [16] examined various aspects of data quality dimensions in different specific decision contexts and have demonstrated the tradeoffs of providing complete data sets. They speculate that under certain circumstances discarding data to make remaining data more consistent may enhance decision making. More specifically, intelligent traffic systems with their variety and velocity of information sources depend heavily on integrated complete information to make proper decisions [17]. Using ontologies that mapped the various pre-processed data sources into single domain space query efficiency was substantially increased. Similarly, crisis responses require information quality and completeness at the right time [18].

We have taken the approach to make information quality, consistency and completeness fully transparent. By using the concept of information supply chains and information dependencies, missing information was put into a context. Consequently, we built a framework which signaled timely and in specific contexts only any incomplete data sets.

III. Query Support in ISCS of an Immuno-genicity Case Study

Immunogenicity has recently drawn much attention as severe adverse effects from the treatment with therapeutical proteins may affect the health of patients [19]. In the best case, antibodies, produced as a result of humoral immune response, do not seem to alter pharmacokinetics of the drug and hence may still exhibit its desired therapeutical effect. In other cases, however, these antibodies may even cross-react with native proteins and start inducing unpredictable severe side effects. However determination of immunogenicity is not straightforward as it is influenced by many known and yet still unknown parameters and this phenomenon poses a threat to predict immunogenicity [19, 20]. Following section introduces a case study to determine potential immunogenicity by demonstrating the feasibility of an ISC framework to tackle timely information availability.

A. Case Study Generation for Experiment Setting

In determining potential immunogenicity of a therapeutic protein, derived conclusions from several assays need to be appropriately integrated as each assay has its limitations. These assays are carried out in different laboratory conditions and hence require a common understanding of its underlying experimental designs, results and its conclusions. We recognize that elicited information traverse domain boundaries and subject to specific semantics and vocabularies used in a particular domain. We attempt to present a case study which involves capturing of this information from two independent workflows. The case study demonstrates that the results of the individual workflows lack constructing a sense, but when integrated provide a meaningful conclusion (Fig. 4). The diagram shows how two workflows are integrated using ISC (integrated information flow is shown in the middle). Integration is achieved by publishing relevant protocol data from each individual workflow (Workflow 1 and 2 in Fig. 4) into the ISC using common agreed terms and vocabulary that are defined in the underlying ontology. The ISC is then

formed from these specific terms and relevant information from these terms integrated, as will be outlined below in more detail.

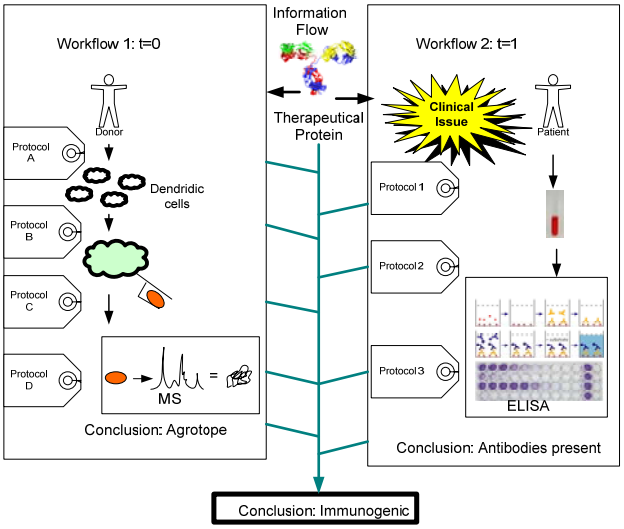

Figure 4. Test cases comprised of two workflows were used for testing feasibility of ISC framework. Resulting information flow is shown in the middle.

In the first workflow, it is tested whether fragments of therapeutical protein are expressed on the surface of specialized cells and thereby potentially attracting the humoral immune system. Protein fragments from specialized cells from donors that are treated with a therapeutical protein are analyzed (Fig. 4, Workflow 1). If protein fragments from the therapeutical protein are detected using sophisticated analytical techniques, the conclusion of this workflow is that the therapeutical protein is potentially immunogenic. Here, the specific terms represent information from the therapeutical protein and all assay protocol related information that was used for analytical procedures and their results. In order to further strengthen the potential immunogenicity a separate workflow – independent from Workflow 1 in Fig. 4 - is introduced. This workflow starts when a clinical issue is reported after a patient has been treated by a therapeutical protein. In this case, the second workflow uses ELISA technique where blood is analyzed to determine the presence of antibodies (Fig. 4, Workflow 2). Like in the first workflow, information from the same specific terms is collected, namely therapeutical protein and all information from assay protocols used for analytical procedures and results. Conclusions from both the workflows, when integrated on the basis of the same terms - in this case the therapeutical protein - enable the researchers to determine the likelihood of possible immunogenicity of the drug. For example, if both workflows have introduced the same therapeutical protein and both conclude a positive result of their independent workflows the finding that this protein may exhibit an immunogenic effect is strengthened significantly.

B. Test Cases and Analysis of the Case Study in the Context of ISC

Worthiness of conclusion, as outlined in the case study above, is bound to have timely access to all relevant provenance data of underlying scientific records. For example, a scientist familiar with a specific assay technology needs different provenance of information than a non-expert in order to have trust in obtained information. Hence, particular data that is considered by one user may not be considered at all by another user.

In order to establish the fact mentioned above, test data were collected from safety immunology laboratories that were involved in the case study along the two workflows as depicted in Fig. 4. In total, 130 information resources were captured, categorized using agreed ontology and annotated to provide provenance information. This pre-processing of information was performed manually. Information included various sources such as extracted summary data, investigations held in PDF or Word, regulations and SOP's shown on PowerPoint presentations, procedures of testing principles, actual data from observations, results from conducting experiments and clinical data from five patients that were made anonymous. In addition, about 30 irrelevant information resources were introduced to test how the ISC framework would deal with such information in real case scenarios.

In order to determine information integration using ISC approach quantitatively, we introduced precision and recall [21] across different information dimensions and in addition information completeness as outlined in section II F to demonstrate an overall performance of the test system. Specific queries were formulated for which adequate answers required integration of information that was spread across different and independent information units as outlined above. Comprehensibility of obtained information for answering such a question required a measure of how complete the retrieved information set was. This experiment revealed of how much information dependencies were satisfied and then the finding was verified by interviewing a focus group. The focus group consisted of three immunology experts.

Measuring trust in obtained information was approached using details of provenance information and observation of scientists how they viewed such information. Finally, visualization of information that is how the graphical user interface displayed information resources was assessed by directly observing scientists how they used the ISC framework.

C. Proposed ISC Framework

The ISC framework was designed to be implemented using RDF/S based on XML notation. Identified information units along the workflows of the case study were mapped into RDF classes and from there subclasses defined. All classes were routed from top class Information Unit, which itself was a RDF class. Main classes were considered immediate subclasses of class Information Unit. For example, class Report, Observation and Project respectively were all main classes. However, these classes had a too wide scope in providing

semantically consistent information as these information containers may have spanned organizational boundaries that did not provide sufficient detailed contextual information necessary to satisfy trusted provenance information. Hence, the more specialized an Information Unit was the more semantically consistent its information source was considered. For example, Protocol defined subclasses such as Guideline, SOP, Regulation etc. and each of these specializations provided specialized contextual information (Fig. 5).

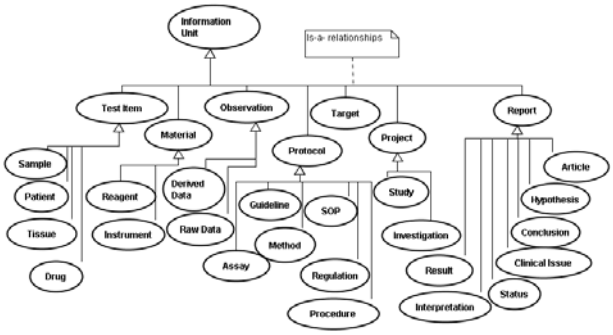

Figure 5. RDF based ontology for high level abstraction of data model used in ISC framework.

The prototype of the ISC framework was organized as an ordered set of specialized Information Units which were connected through concept Information Channel. Information resources were published using Universal Resource Identifiers (URI) into the concept Information Channel. This allowed merging identical resources - published and annotated in different locations that used agreed concepts - and yielded additional information along labeled edges which were not present in the original (unmerged) data set [22]. The Information Channel concept allowed for setting up highly specific queries within a semantic context that were the basis of ISCs; that is, specialized information units – e.g. type of assay used in certain experiments, issues arising in interpreting an immunological profile of a given sample, etc. – could be linked together into an ISC using same underlying concept while providing trust by provenance information through common agreed annotation mechanisms (Fig. 6).

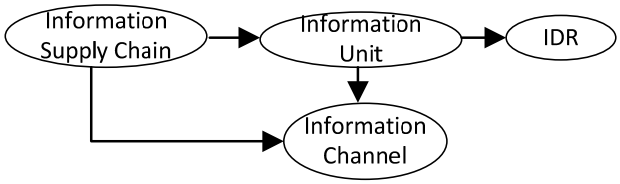

Figure 6. Information Supply Chain Model.

We have modeled information dependency relationship (IDR) using the domain and range concept provided by RDF Schema. The schema allows generating a particular graph with specific information sources dynamically; for example Information Unit of type Clinical Investigation may be dependent on Information Units of type Hypothesis, Issue, Drug and Sample (refer to Fig. 3). For integrating information based on ontology, external RDF ontology description model was used which enabled the researchers easy configuration and adaption of ontology to test feasibility of the framework.

D. Experimental Setup

A specific ISC was configured to collect and integrate relevant information about predictions of immunogenicity of certain therapeutic drugs along the resulting information flow from the two workflows as introduced in section III A (refer to Fig. 4). Information consisted of pre-processed, categorized and published relevant information instances of Protocol, Preparation, Clinical Investigation, Analysis and Determination using agreed terns and ontology as described in section III C. The ISC was setup first without an associated IDR model (Fig. 7). In this setup, a query for determining immunogenicity for example, selected instances of Clinical Investigation directly based on linking immunogenicity as adverse effect to instances of a Therapeutical Protein.

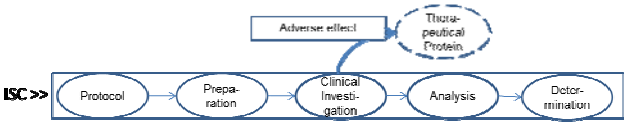

Figure 7. ISC without an IDR Model.

The same ISC was then configured with an underlying IDR model. This was accomplished by associating an IDR model to the respective information used within the ISC (Fig. 8). For example, Clinical Information was supplied with an IDR model as outlined in Fig. 3. In this configuration, the query for determining immunogenicity selected the patient based on all information units which reported immunogenicity as adverse effect and was linked to the therapeutical protein of interest (Fig. 8, IDR of Clinical Investigation).

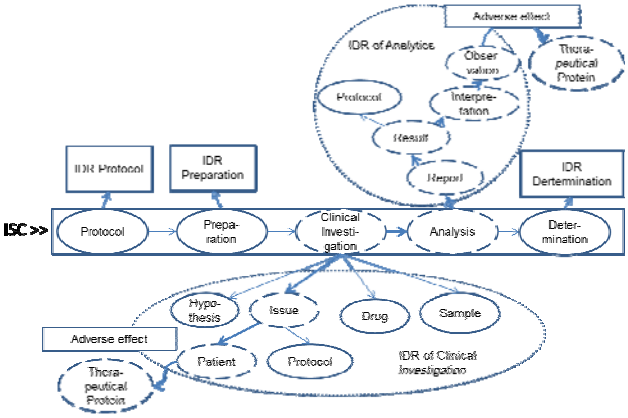

Figure 8. ISC with an underlying IDR Model.

An even more powerful query linking different IDR models within the Information Supply Chain was constructed by the same underlying framework in linking all information units in the ISC with their particular and specific IDR model. Suppose, for example, that information is needed from patient who exhibits an adverse effect due to potential immunogenicity and is confirmed by the same therapeutical protein instance showing a positive result towards immunogenicity testing in Analytics. In order to support such a query, the

therapeutical protein needs to be integrated with all possible instances of several information units that are associated with the term immunogenicity as adverse effect (Fig. 8, linked IDR models of Clinical Investigation and Analytics). Hence an ISC using IDR concept execute queries in a much wider information network than without an IDR model and supports richer integration. This experimental setup enabled us to test timely availability of key information resources to support integration and highly specific queries of two separate workflows as outlined in Fig. 4. Instrumental in providing answers to potential immunogenicity of certain therapeutical proteins of this case study was publishing of the right information from both workflows at the right time. Integration was achieved by using same and agreed ontology in both workflows. Missing information could be easily mapped from the underlying IDR model. For example, if the IDR model requested specific information such as Protocol in a Clinical Investigation (see Fig. 8, IDR of Clinical Investigation) an absence of such information for a particular ISC could be shown and demonstrated to the user. As will be demonstrated in the next section the experimental setup promoted collaboration as missing information could be spotted timely and very specific to the corresponding workflow. Information made available contributed to the precision of a complex query significantly.

E. Results

This section aims highlighting key findings from assessing the feasibility of an ISC framework in supporting complex queries with an indication how well obtained query result can be trusted. In addition missing information - specific to the underlying workflow - could be easily shown, and once made available, contributed to a better query result. This in turn promoted collaboration. The following queries, ranging from basic to rather complex information integration, were executed (Table I) and precision and recall of obtained information was recorded.

TABLE I.
QUERY TYPES

Query	*Search Term (s)*
Search protocol related info about immunogenicity	immunogenicity, protocol
Search patients in which immunogenicity is reported as an adverse effect	immunogenicity, adverse effect, patient
Search information in which immunogenicity is reported as an adverse effect	immunogenicity, adverse effect
Search patients in which immunogenicity is reported as an adverse effect	immunogenicity, adverse effect, patient
Search drugs with potential immunogenicity reported in clinical studies as adverse effects	immunogenicity, clinical study, adverse effect, drug

In addition, Information Completeness (IC) was measured for ISC using IDR approach. General assumptions were that all relevant information was published into appropriate information channel. The search performance is summarized in Table II:

TABLE II.
ISC SEARCH PERFORMANCE USING DIFFERENT QUERY SUPPORT MODELS (P: PRECISION, R: RECALL, IC: INFORMATION COMPLETENESS)

Query	Simple		Semantic		ISC		ISC with IDR IC=0.6		
	P	R	P	R	P	R	P	R	IC
1	0.8	1	--	--	1	0.5	1	0.3	0.6
2	0.8	0.4	0.8	1	1	0.4	1	1	0.5
3	0	0	0.9	1	1	0.7	1	0.5	0.6
4	0	0	0.1	1	1	1	1	1	0.5
5	0	0	0	0	0	0	1	1	0.6

The results showed that executions of ISC based queries performed significantly well in answering more complex questions. The most complex query (Query 5, Table II) could only be answered by ISC using IDR model and hence demonstrated superiority in integration capability over an ISC without an IDR model. The calculation of recall considered all retrievable information without restricting to specific information channel, which otherwise would result in an R value of 1. Calculated IC values indicated that the information content was not exhibiting all necessary details despite the fact the search performed well (i.e. precision showed value of 1). This was in agreement with the overall IC value for the ISC (IC = 0.6), which suggested that demand and supply of information was not in balance. Collaboration was encouraged in supplying particular information that was not present based on information dependencies, as exemplified in Fig. 9a: Four instances of specific information for a particular information instance (in this case, protocol_0005) needed to be supplied if the information set were said to be complete. These dependencies were inferred from underlying IDR model and were displayed on a separate page which listed first required information units (in the case of protocol_0005 Procedure, Target, Result and Regulation) and then information unit instances that satisfied the condition in forming a relation to the root information (here, again protocol_0005) which is demonstrated in Fig. 9b.

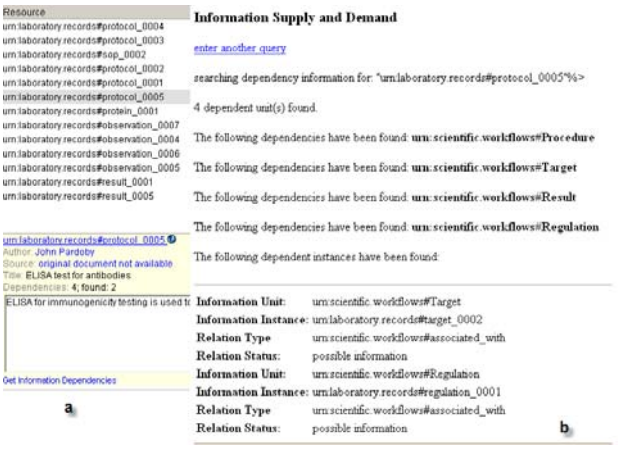

Figure 9a (left) and 9b (right). a) Four information dependencies are indicated for protocol_0005, however two are present. b) Details about the information instance demonstrate the imbalance of required information.

This procedure enabled the scientists to supply all information required for the particular ISC. It animated to browse through the demand and supply chain of information by visualizing all required properties and its values from a particular information instance. Whenever information was indicated to be short on supply (i.e. missing, as indicated in Fig. 9) it was generally appreciated and influenced researchers positively in providing such information. In turn, query precision increased significantly once the right information within the right context was supplied.

Yet, how particular information was dependent on other information was varying among researchers. The underlying IDR model could be easily modified to suit the personal requirements of a scientist. The same information instance, protocol_0005 in Fig. 9b, was accessed with a different IDR model, in which Regulation was not considered relevant for a Protocol. This resulted in same details for the particular information instance (Fig. 10a), yet exhibiting only three information dependencies (Fig. 10b). This was achieved by adjusting the IDR model such that Protocol was no longer dependent on Regulation.

Figure 10a (left) and 10b (right). a) Only three information dependencies are indicated for information instance protocol_0005 after adjusting the IDR model. b) Details about supplied dependencies showed two related information instances.

In order to further test query performance based on the IDR support model, each query was executed on the ISC framework by varying information completeness, ranging from low (IC= 0.1) to a full complete information set (IC = 1.0) with respect to the underlying IDR model. Precision and recall of obtained query result was recorded and are summarized in Table III.

TABLE III
QUERY PERFORMANCE OF ISC FRAMEWORK WITH VARIOUS INFORMATION COMPLETENESS.

Query #	ISC (IC = 0.1)		ISC (IC = 0.6)		ISC (IC=1.0)	
	P	R	P	R	P	R
1	1	1	1	1	1	1
2	1	1	1	1	1	1
3	0	0	1	1	1	1
4	0	0	1	1	1	1
5	0	0	1	1	1	1

The results showed that when information was balanced with regards to demand and supply of information all queries performed very well. IC values from 0.6 up to 1.0 showed identical query performance, whereas lowering the information supply down to an IC value of 10% resulted in poor performance for the more complex queries in terms of precision and recall. Hence, the IDR concept showed high selectivity for queries in that either it would retrieve the right information (e.g. precision is high), or it won't retrieve anything (e.g. recall is low). This result was in agreement with individual assessment of subject matter experts and supports the concept of using IDR in ISCs in performing complex queries. As eluted above, missing information in this framework could be easily spotted and provided. However, at which IC level a complex query can be trusted could not be established; that is queries which run on ISCs with an IC value of 1.0 are not more trustworthy than an IC value of 0.6. The query support – at least as shown in the results – is the same. Yet, the framework allowed for inspecting missing key information and determining whether such information is indeed necessary for full trust or whether such information is considered only of supplementary value.

IV. DISCUSSION AND CONCLUSION

The obtained results showed that an ISC framework using an underlying IDR concept is able to provide query support for answering complex questions, especially when information has to be integrated with complex information dependencies. This is because the underlying IDR model provide for a semantic rich environment to allow for integrating relevant information. It also supports the concept of balanced information with regards to the demand, as pre-scribed in the IDR model, and supply, provided by actual information instances. Information with low IC value indicated low information consistency and was confirmed by a poor query performance (e.g. ISC with an IC of 0.1). Once information content was increased the queries were able to integrate all necessary information instances and to generate the result. The result indicate a kind of "all-or-nothing" behavior with regards to IC values within an ISC framework in that either all necessary information instances are found, or, if one is missing, the respective query will not be supported and will fail. We have also shown how this framework supports the collaborative aspect in providing missing information at the right time and with required specificity. This in turn increased the IC value and hence the information quality of an ISC.

Hence it can be said that these highly specific ISCs provided for a powerful query support for information integration and missing information required for complex query support could be spotted quite easily and made available timely based on the information dependency relations.

We have demonstrated how the concept of information completeness in ISCs supports complex queries simply by supplying information demand of the underlying IDR model. We established a basis how balanced information

demand and supply is able to improve trust in answering complex questions through the use of personalized IDR models. These models provide for collaboration and rich information integration on the basis of a semantic consistent context. The level of consistency can be computed based on how complete particular information set with regards to its IDR model is. Such consistency provides a high level of trust. However, we still need to enhance the test data set and allow for more than one information channel into which information can be published. This would also allow establishing a link between IC values and query performance, e.g. how trustworthy a particular result from a complex query indeed is.

REFERENCES

[1] A. Anjum, P. Bloodsworth, A. Branson, T. Hauer, R. McClatchey, K. Munir, D. Rogulin, J. Shamdasani, "The Requirements for Ontologies in Medical Data Integration: A Case Study", 11th International Database Engineering and Application Symposium, IDEAS, 2007, pp. 308-314.

[2] H. I. Abas, M. M. Yusof, S. A. M. Moah, " The application of ontology in a clinical decision support system for acute postoperative pain management", Semantic Technology and Information Retrieval (STAIR), 2011, pp. 106-112

[3] K. Mecheri, M. Boufaida and L. Souici-Meslati, "Data mediation towards semantic web service interoperability". Int. J. of Web Science, vol. 1(3), 2012, pp. 163-179.

[4] R. Bose, and J. Frew, "Lineage Retrieval for Scientific Data Processing: A Survey". ACM Computing Surveys, vol. 37(1),2005, pp. 1-28.

[5] D. De Roure and C. Goble, " myExperiment – A Web 2.0 Virtual Research Environment", International Workshop on Virtual Research Environments and Collaborative Work Environments, Edinburgh, UK, 2007.

[6] R. A. Becker and J. M. Chambers, "Auditing of data analyses", J. Sci. Stat. Comput., vol. 9(4), 1988, pp. 747–760.

[7] M. Del Pillar Angeles and L. M. MacKinnon, " Managing data quality of integrated data with known provenance", Int.J. of Information Quality, vol. 2 (3), 2011, pp. 244-263.

[8] A. Polonsky, A. Six, M. Kotelnikov, V. Polonksy, R. Polly and P. Brey, " Semantic Laboratory Notebook: Managing Biomedical Research Notes and Experimental Data", Proceedings of the European Semantic Web Conference (ESWC 2006).

[9] K. N. M Boulos and T. Wheelert, "The emerging Web 2.0 social software: an enabling suite of sociable technologies in health and health care education", Health Information and Libraries Journal, vol. 24, 2007, pp. 2-23.

[10] C-C Lin and C-W. Lin, "Defective item inventory model with remanufacturing or replenishing in an integrated supply chain", Int. J. of Integrated Supply Management, vol. 6(3/4), 2011, pp. 254-269.

[11] A. Saito, K. Umemoto and M. Ikeda, "A strategy-based ontology of knowledge management technologies", Journal of Knowledge Management, vol. 11(1), 2007, pp. 97-114.

[12] L. Kerschberg, and H. Jeong, "Just-in-Time Knowledge Management". Keynote Talk, Third Conference on Professional Knowledge Management, April 10-13, 2005, Kaiserslautern, Germany.

[13] S. Sun and J. Yen, "Information Supply Chain: Unified Framework for Information-Sharing", Proceedings of IEEE International Conference on Intelligence and Security Informatics, 2005.

[14] Y. L. Simmhan, B. Plale, and D. Gannon,"A Survey of Data Provenance in e-Science", SIGMOD Record, vol. 34(3), 2005.

[15] P. Walsh, and A. Koumpis, A. "Introducing the concept of information supply chains: the Buddy project", Logistic Information Management, vol. 11(2), pp. 74-79, MCB University Press, ISBN 0957-6053, 1998.

[16] D. P. Ballou and H. L. Pazer, "Modeling Completeness versus Consistency Tradeoffs in Information Decision Contexts", IEEE Transactions on Knowledge and Data Engineering, vol. 15(1), 2003, pp. 240-243.

[17] W. Yang and Q. Zhu, "Ontology-based semantic fusion of traffic information", International Conference on Computer Science & Service System (CSSS), August 2012, pp. 769-772.

[18] R. A. Gonzales and N. Bharosa, "A Framework Linking Information Quality Dimensions and Coordination Challenges during Interagency Crisis Response", Proceedings of the 42nd Hawaii International Conference on System Sciences, 2009.

[19] F. Bader, "Immunogenicity of Therapeutic Proteins: A Case Report. Scientific Considerations Related to Developing Follow-On Protein Products", FDA Public Workshop, September 14 & 15, 2004.

[20] M. Barbosa and E. Celis, "Immunogenicity of protein therapeutics and the interplay between tolerance and antibody responses", Drug Discovery Today, vol. 12(15-16), 2007, pp. 674-681.

[21] S. M. Shafi, and R. A. Rather, "Precision and Recall of Five Search Engines for Retrieval of Scholarly Information in the Field of Biotechnology", vol.2 (2/12), 2005.

[22] D. Brickley and R. V. Guha, " RDF Vocabulary Description Language 1.0: RDF Schema", W3C Recommendation 10 February, 2004.

ACKNOWLEDGMENT

The authors wish to thank Harald Kropshofer and Guenther Fischer for their valuable feed-back and acting as focus group.