An integration-based conceptual framework for scientific information analysis

Wei Chen†, Rui Yang, Tao He, Shuo Wang

Wuhan Branch of National Science Library, Chinese Academy of Sciences,
Wuhan, Hubei 430071, China

*E-mail: chenw@whlib.ac.cn
www.whlib.cas.cn

Jiangping Chen

Department of Information Science, University of North Texas
Denton, Texas 76203, USA
E-mail: jiangping.chen@unt.edu

The era of big data brings both opportunities and challenges to scientific information analysis (SIA) and intelligent information services performed at the National Science Library, Chinese Academy of Sciences (NSLC). We are in urgent need of developing a new SIA framework to expedite big data acquisition and processing, and to improve the quality of information services. This paper describes the traditional SIA workflow currently applied at NSLC with a case study. It also reviews progresses on massive heterogeneous data integration, data management and analytics methods, and their applications. We then propose an integration-based conceptual framework for SIA through an examination of the limitations of current workflow. The new framework is characterized with the development of a Knowledge Resources Integration System (KRIS) that can store, organize, process, and visualize heterogeneous data. We explain the functions and characteristics of the proposed framework and strategies to implement a web-based data warehouse system based on it. The paper concludes with a discussion of future research to implement KRIS for SIA.

Keywords: Scientific Information Analysis; Conceptual Framework; Heterogeneous Data; Data Integration; Data Analytics.

1. Introduction

Scientific Information Analysis (SIA) is one of the core knowledge services conducted at the National Science Library, Chinese Academy of Sciences.
(NSLC). The procedures of SIA consist of searching for, collecting, processing, and interpreting information in order to gain an awareness of state-of-the-art and to forecast potentially important future developments in specific technology areas for science policy decision-makers. Specifically, SIA includes trend analysis, emerging topic detection, competition and collaboration analysis, and evidence-based strategy and policy analysis. SIA at NSLC has contributed to the decision-making of Chinese Academy of Sciences (CAS) and has been continually supported by the organization.

Current SIA at NSLC follows a framework which consists of six sequential yet iterative phases: Planning, Information collecting, Information processing, Information analysis, Results documentation, and Evaluation and feedback. The SIA task at each phase is mainly conducted manually by information analysts. One of the biggest problems of this framework is that it requires considerable time and laborious effort to complete each phase, especially the information collecting, integrating and analysis phases are time and labor intensive. Consequently, the productivity of SIA following this framework has been negatively affected.

As big data offers unprecedented opportunities for not only accelerating scientific advances, but also enabling new modes of discovery (Honavar, 2014), scientific research is undergoing a data-intensive paradigm shift (Hey, Tansley & Tolle, 2009). SIA, which serves scientific decision-making, is demanded to provide timely, accurate, and comprehensive results through knowledge analysis and knowledge discovery (Zhang, 2012). To meet the challenge, NSLC is in urgent need of developing a new SIA framework to expedite big data acquisition and processing, and to improve information service quality.

A well-organized data management system integrated with a series of proper analytical methods may help the SIA tasks to be conducted more effectively. The purpose of this study therefore is to redesign the SIA framework by adding integrated data management and analysis capabilities into the phases. The new integration-based SIA framework is expected to facilitate fast information collecting and analysis. It will enable information analysts to quickly access required information and conduct in-depth intelligence analysis through a variety of analytical methods.

This paper is organized as follows: It first describes the traditional SIA framework or workflow applied at NSLC with a case study as well as the workflow’s problems and limitations. Then it reviews existing literature on massive heterogeneous data integration, data management and analytics methods, and their applications. Next it proposes an integration-based conceptual framework for SIA, which characterized with the development of a
Knowledge Resources Integration System (KRIS) that can store, organize, process, and visualize heterogeneous data. Finally, several future research directions to implement KRIS for SIA are discussed.

2. Traditional SIA Workflow and A Case Study

2.1. Traditional SIA workflow

Scientific decision-making is a process of information convergence. To facilitate it, NSLC has carried out a wide range of exploratory efforts, and has established an intelligence service model in line with decision-makers’ requirements. Current SIA workflow at NSLC is roughly classified into six sequential yet iterative phases, which are illustrated in Fig. 1.

(i) Planning: In this first phase, we analyze the requirements from decision-makers to identify analysis subjects, technology areas and goals, and to design an analytical plan in accordance with questions-oriented approach. In general we are asked to investigate a specific technology policy landscape, and to provide technology development status and trends, competitiveness assessment, and proposed measures and recommendations.

(ii) Information Collecting: According to the analytical plan, various types of relevant information are scanned manually from different information sources. Through pre-screening and information validation, the relevant information is collected as resource collections, which are usually stored in decentralized personal repositories.

(iii) Information Processing: This phase involves data classification, metadata extraction, data cleansing, data normalization, and preservation. Thomson Data Analyzer™ (TDA), a commercial analysis software designed to work with all types of structured content like patent and scientific literature databases (Clarivate Analytics, 2017a), is used to process raw structured data. Unstructured data, such as policy and report documents, have to be handled manually due to the lack of appropriate techniques and tools.

(iv) Information Analysis: In this phase, we apply both qualitative and quantitative analysis techniques to synthesize the data and discover new knowledge. Quantitative methods are mainly limited to the bibliometric analysis at present, which analyzes scientific literature and patents to discover and evaluate competitors and collaborators. Qualitative methods, such as the Delphi method, literature review, focus group discussion, and SWOT analysis are used to analyze and validate text data.
(v) Results Documentation: In this phase, the analytical results are documented and various intelligence products are generated. Depending on the needs of decision-makers and reporting requirements, intelligence products take many forms such as newsletters, digest compilations, briefings for decision-makers, analysis reports, presentations, journal papers, and commentaries.

(vi) Evaluation and Feedback: After analytical products are completed, information analysts commit a self-assessment and solicit feedback from domain experts or the users. The satisfactory outcomes are accepted by the decision-makers who would put them into practice, or consider them for decision-making, whereas unsatisfactory ones are refined by going through previous phases again based on the feedback. Sometimes, decision-makers put forward new requirements in response to the changes in practice or emerging situations, and then ask for new analyses.

![Diagagram](image)

Fig. 1. Traditional scientific information analysis workflow at NSLC

2.2. A case study

Below we present a case study to demonstrate the actual operations using current SIA workflow. The case: The decision-making body of CAS requested an investigation of global status of energy storage technologies, and recommendations for domestic development within two months.

2.2.1. Planning

In responding to this request, a task force comprised of three energy information analysts was formed and led by a team leader. The task force analyzed the
requirements at first, thereby identifying that the analysis subject was energy storage, the technology areas included pumped hydro, compressed air energy storage, batteries, to name a few, and the goals were to investigate global status and to provide recommendations. Then the task force reframed the request in the form of four important questions, and designed the analytical plan to address these questions, as shown in Table 1. The related work from information collecting to documentation on each question in one or more technology areas was assigned to each information analyst. The design of analytical plan and the arrangement of the project took about 3 working days.

<table>
<thead>
<tr>
<th>Defined Questions</th>
<th>Analytical Strategies</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. What do major national governments do to promote energy storage technology development?</td>
<td>Describe energy storage technology policy landscape, including ongoing and planned R&amp;D strategies, programs, public investment in major countries.</td>
</tr>
<tr>
<td>2. What are the status of energy storage technologies, and are the trends of R&amp;D?</td>
<td>Assess technological state-of-the-art and anticipated development trends.</td>
</tr>
<tr>
<td>3. Who are the main rivals and potential partners?</td>
<td>Conduct competitive analysis and benchmarking, including comparison and evaluation of major research institutes, key industrial players, and leading scientists and engineers.</td>
</tr>
<tr>
<td>4. What R&amp;D plans should China implement in energy storage areas to bridge the gap with leading nations?</td>
<td>Develop recommended strategies and practical solutions.</td>
</tr>
</tbody>
</table>

2.2.2. Information collecting

During this phase, information analysts first searched for documents and numerical data files related to the defined questions from open network resources, subscribed databases, and internal data files and reports. Then they collected and classified these materials into four types of collections, and stored them in decentralized personal repositories:

(i) Technology policy documents collection, such as energy storage R&D strategies, roadmaps, program plans, R&D budgets, and projects of major national governments (e.g., US, Europe, Japan, China). This collection contained 34 documents.

(ii) Technical and industry reports collection, including the publications and review papers from various energy policy research institutes, academic institutions, consulting agencies, independent think tanks, and industry associations. Some of these reports reviewed the development and prospect of energy storage technologies, and some discussed market status and potential. This collection contained 164 documents.
(iii) Scientific literature collection, in which research articles were searched and downloaded from Web of Science®. This collection contained 45,257 articles.

(iv) Patent collection, in which patents were searched and downloaded from Derwent Innovations Index®. This collection contained 44,402 patents.

Due to the diversity of information types, wide range of sources and complexities of data searching, the task force spent nearly 20 days gathering and screening relevant information in this phase. Although the manual searching and screening of relevant materials brought high precision on the retrieval, it is difficult to predict the completeness of the collections.

2.2.3. Information processing and analysis

In this phase, the task force took different analytical methods for different types of data. Since bibliometric analysis were usually used to assess competitiveness and research performance of institutions and individuals, information analysts utilized Excel, TDA, and Thomson Innovation® which is a commercial platform with built-in patent search, analysis and visualization tools (Clarivate Analytics, 2017b), as primary tools to process and analyze the research papers and patents, including measure of scientific output by paper counts, evaluation of research impact by citation analysis, and identification of hot and emerging research topics by cluster analysis. The analytical results partially addressed the second and third defined questions specified in Table 1.

On the other hand, information analysts made use of literature review and focus group discussion to analyze policy and report documents. The focus group, consisting of all of the energy information analysts, discussed everyone’s review regularly and synthesized disparate pieces of information to identify correlated intelligence and patterns related to different technology areas, then to interpret newly developed knowledge, such as the comparison of similarities and differences of R&D strategies among major nations, the effectiveness of their policies, the status of various technologies, and the market potential. Next, the task force designed expert questionnaire based on the above preliminary results, utilized the Delphi method to seek the opinions of specialists in a variety of energy storage technology areas, and incorporated their forecasts on technology trends into the information analysis. The analytical results addressed the first and second defined question specified in Table 1.

As the core component of this SIA task, this phase took the longest time to complete, nearly 25 days. The main reason for taking so long was that most of the analysis completely depended on human integration and interpretation of
large volumes of heterogeneous data, which is a time-consuming and labor-intensive process.

2.2.4. Results documentation

In order to disseminate the analytical results and new discovery to support decision-making, the task force spent almost 10 days preparing three types of intelligence products, including:

(i) A summary for decision-makers, which was a summary of the full report intended to aid decision-makers on the energy storage issue. It contained a high-level description of main results and, more importantly, policy and technical recommendations for future development.

(ii) A complete analysis report, which included the detailed analysis process and findings. It was organized into six chapters: introduction, methodology, energy storage technology policy landscapes in major nations, status and trends of energy storage technologies, comparison and evaluation of key players, and conclusions and recommendations.

(iii) Formal presentation slides, which focused on key findings illustrated concisely and clearly through graphics, tables or whatever descriptive approaches to facilitate direct communication with decision-makers.

2.2.5. Evaluation and feedback

For the purpose of examining the accuracy and balance of intelligence products, the task force held two review meetings with domain experts and decision-makers of CAS after self-assessment. According to the feedback from reviewers, the main problem of report is lack of deep content mining to discover more policy and technical details as well as sufficient data to support forecasting on technology trends. Because of that, the task force took 6 more days to do additional search and extensive human interpretation of documents, thereby failed to meet the task deadline. Besides, the decision-makers raised concerns about compressed air energy storage technology, and required that the task force conduct further in-depth investigation. Thus, production of intelligence generates more requirements in this iterative process.

2.3. Problems and limitations

In the era of big data, the needs for real-time analyzing and digging deeper into multi-source heterogeneous data are increasing rapidly for decision-making. The volume and variety of data have far outstripped the capacity of manual analysis (Provost & Fawcett, 2013). As analysis tasks often have fixed deadlines, it requires well-organized knowledge management and proper analytical methods
designed to deliver the right intelligence at the right time to produce effective
decision-making (Lee et al., 2012). Apparently current human-driven SIA
approaches have several problems and limitations that cannot accommodate the
changing environment:
(i) Time-consuming. Several phases require considerable time and laborious
efforts, especially in information collecting, integrating, and analysis. These
tasks are still heavily relied on manual gathering, processing, integrating,
and interpreting a large set of documents.
(ii) Limited knowledge discovery capabilities. Due to the fact that the majority
of the data is in the forms of multi-attribute and unstructured form of
documents, information analysts are still only able to analyze a small
percentage of the data acquired and stored.
(iii) Data management and sharing problems. The policy and reports datasets
usually stored in decentralized personal repositories without proper
infrastructure to share and integrate the data, which may not be well
managed and used effectively.
(iv) Methodological problems. Most of the analyses have been descriptive,
small-scale studies lacking theoretical framework and quantitative content
analysis methodologies and techniques.

This intelligence service model of our SIA is a type of descriptive
information analytics which looks at past performance and understands that
performance through mining historical data to look for the reasons behind past
success or failure. It lacks predictive power. Although we still need descriptive
information analytics to know what really happened in the past, there has been
much more requirements for integrating and analyzing multi-source
heterogeneous data to extract predictive patterns, also referred as predictive
analytics. Predictive analytics can optimize limited resources to make better
decisions and take actions for the future in the context of big data. However,
current SIA, limited by the manual data collecting and analysis approaches,
cannot carry out predictive analytics.

3. Opportunities in Big Data Analytics: Related Studies

Current SIA workflow can be improved by integrating big data approaches and
technologies. The main features of big data (volume, velocity, variety,
variability, value, etc.) are creating new challenges for data management and
analysis (IDC, 2011; NIST, 2015; IBM, 2017). However, big data has attracted
tremendous interests from industry, government and research community, and
created high expectations for positive outcomes (Power, 2014). New discoveries
in a wide range of disciplines are increasingly being driven by the ability to
acquire, share, integrate, analyze, and build predictive models from data (Honavar, 2014). The big data paradigm promises to turn imperfect, complex, and often unstructured data into actionable information, delivering a prospect for cost-effective opportunity for improved decision-making in critical areas such as scientific research, business intelligence, healthcare, public administration, and national security (Provost & Fawcett, 2013; Hilbert, 2016).

Typically, information analysts increasingly need to integrate data from multiple sources, and of multiple types, into their strategic data analyses. The variety of heterogeneous data being used, rather than the scale of the data being analyzed, is the limiting factor in data analysis efforts (Hendler, 2014). To address these challenges, the research community and industry have proposed various solutions for big data systems.

3.1. **Data integration**

The integration of diverse data sources and novel ways drives our ability to see new things and develop predictive analytics. As traditional approaches of integration are inefficient in handling big data situation, exploring how to use the data association and integration to maximize the value of big data has been a hot research topic, especially when the deeper integration remains difficult. Besides dozens of expensive commercial data integration platforms (Gartner, 2016), a variety of user-friendly and feature-packed open source data integration tools have flourished recently (Walsh, Rodrigue & Mummadi, 2016; Hassani, 2017). Many of them like Kettle (Pentaho Corporation, 2017) and Talend Open Studio (Talend, 2017) have intuitive graphical user interfaces with drag-and-drop functionality for ease of usage, run on any platform/operating system, and can be deployed in a variety of configuration options. These effective, no-cost solutions can be explored to develop various economical big data applications.

Another possible tool is DeepDive (Zhang, et al., 2017), an open source data management system that is able to create structured data (SQL tables) from unstructured information (text documents) and integrate such data with an existing structured database. DeepDive leverages the effectiveness and efficiency of statistical inference and machine learning for difficult extraction tasks. It has been used in a number of domains from pharmacogenomics to paleobiology to antihuman trafficking enforcement.

Other researchers have also contributed a lot in this field. Chen (2017) developed a data infrastructure for the Intelligence and Security informatics (ISI) community, with primarily focusing on data collection, data management, and data access. The infrastructure consists of online archives and analysis tools, by
which integrating a wide array of open source data allows the security research community to more easily collaborate with other members of the community. Ma, et al. (2017) proposed a data integration framework based on Unified Concept Model (UCM) to address real-world gasoline and natural gas safety supervision problems. By following the structure of UCM, data from different sources is automatically transformed into instance data, stored in a graph database, and linked together by using semantic similarity computation metrics. Daraio, et al. (2016) proposed an Ontology-Based Data Management (OBDM) approach to integrate heterogeneous data sources, including big scholarly data (such as publications and citations) to support the assessment of research and develop “science of science” policy models. Meng, et al. (2016) created ScholarSpace (C-DBLP), an author-centered Chinese literature integration system in Computer Science to support facet-based academic information retrieval of categories such as scholars, research interests, and research topics. Williams, et al. (2014) provided a case study of CiteSeerχ, a digital library and search engine for academic documents. It integrates data from across the web and performs automatic extraction, clustering, entity linking and name disambiguation on data.

3.2. Data analytics

Data analytics is the final and most important stage in the value chain of big data, with the purpose of extracting useful values, suggesting conclusions and/or supporting decision-making. Generally, data analysis techniques can be categorized into six types: structured data analytics, text analytics, web analytics, multimedia analytics, network analytics, and mobile analytics (Hu, Wen, Chua & Li, 2014). The vast majority of data analysis techniques belong to either descriptive analytics or predictive analytics, the latter of which has received increasing attention in decision-making process in recent years. According to Wlodarczyk and Hacker’s quantitative analysis results (2014), the trends in combined use of big data and predictive analytics technologies demonstrate that big data is the driving force behind predictive analytics.

Many tools for big data mining and analysis are available online, including expensive commercial software/platform, as well as open source tools such as Rapidminer, Weka, and KNIME (Yaqoob et al., 2016), most of which are Java-based and platform-independent. However, data analysis is a broad area, which frequently changes and is extremely complex. The time and space complexity of data analysis algorithms differ significantly from each other according to different data characteristics and application requirements (Chen, Mao & Liu,
While researchers have created a variety of frameworks to treat problems of extracting useful knowledge from data, they are usually confined to limited data types or certain application scenes. Therefore, current data integration and analytics platforms need to be carefully examined and tested before they can be adapted or integrated into SIA workflow. They may be not able of solving the challenges of analyzing multidisciplinary, dynamic, and complex data without customization.

4. An Integration-based Conceptual Framework

We would like to propose a modified SIA conceptual framework to address the problems and limitations of current SIA workflow. This integration-based conceptual framework mainly focuses on reforming the search and analysis phases based on big data management and analytical methods. The new framework is characterized with the development of the KRIS that can access, store, retrieve, organize, process, and visualize heterogeneous data, including utilizing new technologies to link data across datasets, and integrate and synthesize structured and unstructured data to gain further insights.

4.1. Design principle

In principle, KRIS design should consider functionalities, flexibilities, and usability in big data environment for collaborative work. The system is designed to collect a wide variety of R&D inputs and outputs information, including R&D policies, funding and projects, institutions and professionals, scientific literature, patents, analysis reports, media news, statistics, and other data. It should offer more accurate analysis and visualization of new patterns and trends in specific research areas. It is also flexible so that different resource specifications, data models, and algorithms can be adapted for information resources discovery, selection, organization, and analytics. KRIS should enable information analysts to quickly access required information and conduct in-depth intelligence analysis through a variety of analytical methods.

4.2. System architecture

Through adoption of big data frameworks and tools to design and build an overall solution from data acquisition to data storage, processing, retrieval, and analytics, KRIS can effectively integrate massive heterogeneous data, perform high-performance computing analysis, and facilitate intelligence services, as illustrated in Fig. 2. Specifically, KRIS consists of five components as indicated
in the “Integration System” layer of the overall architecture. These components are discussed in detail in following sections.

4.2.1. Data acquisition

The first component of KRIS is data acquisition. It should be able to collect distributive, diversified heterogeneous data based on different data acquisition solutions, as illustrated in Fig. 3. In particular, it is proposed to use the big data acquisition framework Apache Flume (The Apache Software Foundation, 2017a) to design a unified data acquisition interface, and develop and deploy different agents and data source middleware, thus collecting the structured data, semi-structured data, and unstructured data.
4.2.2. Data extraction and integration

This component may include tasks such as data cleansing, extraction of entities and relations, aggregation and correlation, validity checks, and other processing, as illustrated in Fig. 4. Converting unstructured or semi-structured data to structured data is mainly carried out by two modules: natural language processing and information extraction. The natural language processing module is designed to perform tasks such as named entity recognition, part-of-speech tagging, syntactic parsing, and coreference resolution. On the basis of that, the information extraction module applies one or more strategies to extract relevant contents from the semi-structured or unstructured texts. Since it is difficult to obtain the hand annotated corpus in practical operation, the semi-supervised extraction modes are mostly used. To prevent semantic drift, the expert intelligence is introduced into the information extraction process to continuously clean the drifting errors.

Fig. 4. Schematic diagram of data extraction and integration

4.2.3. Data storage

The data storage component of KRIS applies open source big data tools such as Hadoop, Hbase, MongoDB to implement data reading/writing and storage. This component should be designed to effectively support the distributed storage issues of the system. Special attention should be paid to high concurrent access to massive data. As illustrated in Fig. 5, this component will be able to perform the following functions:

(i) Storage and management of massive relational data based upon Impala clusters;

(ii) Storage and management of massive unstructured/semi-structured data based upon HBase clusters.
The data storage component should have a shared-nothing architecture so that each node in the system is independent and self-sufficient. Zookeeper will be used to achieve coordination and free horizontal scaling.

4.2.4. Data retrieval

Next component is data retrieval where information can be retrieved through semantic indexing and/or full-text indexing to support quick query, discovery, and computing of information. Fig. 6 illustrates the two types of retrieval approaches. To implement them, Apache Solr mechanisms such as SolrCloud and Solr sharding (The Apache Software Foundation, 2017b) can be used to build a distributed big data retrieval platform instead of using conventional stand-alone search mode. As for the semantic retrieval, semantic repository Virtuoso can be leveraged to index semantic data and to support quick query, discovery and computing of information.
4.2.5. Data analytics

KRIS will also support deeper data analysis beyond data retrieval for the purpose of knowledge discovery and prediction. The data analytics component of KRIS can be built on Spark framework to perform comprehensive analysis and computation on collected data. Taking advantage of this platform, information analysts may be able to produce various intelligence products that can address specific problems.

In summary, with the revised SIA conceptual framework centered around the KRIS, multiple services or products can be further developed. One of them can be an anomaly alert system that can carry out real-time monitoring on various data sources, including government websites, funding agencies, research institutions, industries, news media, statistical databases, scientific literature, patents, etc. When new relevant information emerges or the data anomalies are detected, the alert system will send early warning messages to remind information analysts so that they can conduct analysis in a timely manner.

The second service can be a question-answering system in specific domain based on semantic indexing. This system will provide exact answers to particular questions, which release information analysts from reading every returned items to locate answers after retrieval.

The third service can be an automated report generation system. Regular scientific intelligence reports can be abstracted to form report templates with fixed formats and specific data requirements. In other words, it is possible to automatically extract and compute the relevant data from KRIS, and generate draft reports in accordance with predefined criteria and templates by the assistance of natural language processing techniques. On the basis of that, information analysts can polish the reports by providing valuable insights gained through interacting with domain experts.

5. Conclusions and Future Work

This study proposed an integration-based conceptual framework for SIA, with the development of KRIS as digital infrastructure that can access, store, retrieve, organize, process, and visualize heterogeneous data. The KRIS, if it were implemented, would enable information analysts to acquire and manage relevant data and information and conduct in-depth intelligence analysis through a variety of analytical methods.

Our future research is to refine the design as proposed in this study, implement the system, and evaluate it using real-world decision-making related
problems. Specifically, we would like to carry out the following activities to extend our current study:

Refine KRIS design by identifying differentiated use cases and needs related to analyzing a great amount of heterogeneous data. For a long time, SIA workflow has tried to shoehorn all types of users into a single information architecture. As the paradigm of scientific research has shifted since the rise of big data, the requests from decision-makers will be different from time to time. The new use cases gain importance to explore and clearly document what data needs to be used and how to collect and analyze for decision support. KRIS as a framework is supposed to support different tasks. It is necessary to build or use different models based on different requests. That would be helpful to prepare professionals to manage and analyze a wide array of data, and eventually to develop big data analytics methods and development environments applicable to different domains and for different needs.

Assessing existing data integration and analytics techniques and open source tools so that KRIS could learn from their strengths and avoid the weaknesses. KRIS can be built by integrating multiple existing data integration and analytics tools, which could indeed lead to beneficial outcomes. However, they may also lead to misuse and misanalysis on decision-making. Therefore, current data integration and knowledge discovery tools should be thoroughly investigated to understand their applicability of being adapted for SIA.

Exploring automatic approaches to construct domain-specific knowledge graphs. Knowledge graphs can enhance the effectiveness of traditional information processing task (e.g., information extraction, search, recommendation, question answering) by providing a valuable background domain knowledge (Gomez-Perez, Pan, Vetere & Wu, 2017). Thus, it will be of great significance in SIA. In respond to different requirements from decision-makers, KRIS may need to enrich its domain-specific knowledge graphs from large text corpora with the assistance of domain experts, so that it will improve the efficiency and quality of knowledge services.

Acknowledgments

We thank Dr. Sian Brannon, Mark Phillips, and Dr. Daniel Alemneh for ongoing discussions and their comments. This work is supported by the Youth Innovation Promotion Association of Chinese Academy of Sciences (Grant No. 2017221), and the Documentation and Information Capacity Building Program of Chinese Academy of Sciences (Grant No. Y62G74).
References

document-mining techniques and tools for technology intelligence:
discovering knowledge from technical documents. *International Journal of
Technology Management, 60*(1/2), 130-156.
Integration Framework Based on Unified Concept Model. *IEEE Access, 5*,
5713-5722.
Interoperability Framework: Volume 1, Definitions*. Retrieved from
and Knowledge Graphs in Large Organisations*. Springer, Cham. Retrieved from
https://link.springer.com/book/10.1007/978-3-319-45654-6
http://community.pentaho.com/projects/data-integration/
of Decision Systems, 23*(2), 222-228.
https://www.talend.com/download/talend-open-studio/#4
Walsh, C., Rodrigue, B., & Mummadi, Y. (2016). *Data and Analytics: Open
Source Data Integration Tool Comparison*. Retrieved from
https://www.excella.com/wp-content/uploads/2016/03/Open-Source-DI-
Tool-Comparison_March2016.pdf
big data information extraction and integration in the CiteSeerX digital
library. In *2014 IEEE 30th International Conference on Data Engineering
(ICDE)*. Chicago, IL: IEEE Computer Society.
Wlodarczyk, T., & Hacker, T. (2014). Current trends in predictive analytics of
Yaqoob, I., Hashem, I. A. T., Gani, A., Mokhtar, S., Ahmed, E., Anuar, N. B., &
Vasilakos, A. V. (2016). Big data: From beginning to future. *International
Journal of Information Management, 36*(6), 1231-1247.
DeepDive: Declarative Knowledge Base Construction. *Communications of
the ACM, 60*(5), 93-102.
China Society for Scientific and Technical Information, 31*(8), 788-797.