Developing a Brain Informatics Provenance Model

Han Zhong[1], Jianhui Chen[2], Taihei Kotake[4], Jian Han[1], Ning Zhong[1,3,4], and Zhisheng Huang[5]

[1] International WIC Institute, Beijing University of Technology
Beijing 100024, China
[2] Department of Computing, The Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong, China
[3] Beijing Key Laboratory of MRI and Brain Informatics, Beijing, China
[4] Department of Life Science and Informatics, Maebashi Institute of Technology
Maebashi-City 371-0816, Japan
[5] Knowledge Representation and Reasoning Group, Vrije University Amsterdam
1081 HV Amsterdam the Netherlands
{z.h0912,hanjian0204}@emails.bjut.edu.cn, csjchen@comp.polyu.edu.hk, kotake@maebashi-it.org, zhong@maebashi-it.ac.jp, huang@cs.vu.nl

Abstract. Integrating brain big data is an important issue of the systematic Brain Informatics study. Provenances provide a practical approach to realize the information-level data integration. However, the existing neuroimaging provenances focus on describing experimental conditions and analytical processes, and cannot meet the requirement of integrating brain big data. This paper puts forward a provenance model of brain data, in which model elements are identified and defined by extending the Open Provenance Model. A case study is also described to demonstrate significance and usefulness of the proposed model. Such a provenance model facilitates more accurate modeling of brain data, including data creation and data processing for integrating various primitive brain data, brain data related information during the systematic Brain Informatics study.

1 Introduction

Brain Informatics (BI) is an interdisciplinary field among computing science, cognitive science and neuroscience [15]. It carries out a systematic study on human information processing mechanism from both macro and micro points of view by cooperatively using experimental cognitive neuroscience and Web Intelligence centric advanced information technologies [14]. BI can be regarded as brain science in IT age and characterized by two aspects: systematic brain study from informatics perspective and brain study supported by WI-specific information technologies. A systematic BI methodology has been proposed, including four issues: systematic investigations for complex brain science problems, systematic experimental design, systematic data management and systematic data analysis/simulation [5,16].

K. Imamura et al. (Eds.): BHI 2013, LNAI 8211, pp. 439–449, 2013.

Systematic brain data management is a core issue of the systematic BI methodology. Systematic investigations and systematic experimental design have resulted in a brain big data, including various primitive brain data, brain data related information, such as extracted data characteristics, Related domain knowledge, *etc.*, which come from different research groups and include multi-aspect and multi-level relationships among various brain data sources [9]. It is necessary to realize systematic brain data management whose key problem is to effectively integrate multi-mode and closely-related brain big data for meeting various requirements coming from different aspects of the systematic BI study [16]. Brain data provenances provide a practical approach to realize the information-level (i.e. metadata-level) integration of brain big data. However, the existing neuroimaging provenances focus on data sharing and automatic data analysis, and cannot meet requirements of systematic brain data management. The systematic BI study needs to construct BI-specific provenances of brain data, i.e. BI provenances [16].

In this paper we put forward a provenance model for constructing BI provenances. The remainder of this paper is organized as follows. Section 2 discusses background and related work. Sections 3 and 4 describe such a provenance model and its conceptual framework, respectively. Section 5 provides a case study in thinking-centric systematic investigation. Finally, Section 6 gives concluding remarks.

2 Background and Related Work

Provenance information describes the origins and the history of data in its life cycle and has been studied based on the relational database, XML, *etc* [1,3,8]. In brain science, the metadata describing the origin and subsequent processing of biological images is often referred to as "provenance". For example, Allan J. MacKenzie-Graham et al. divided neuroimaging provenances into data provenances, executable provenances and workflow provenances [12]. However, the existing neuroimaging provenances mainly focus on describing experimental conditions (e.g., parameters of devices and subject information) and analytical processes for data sharing and automatic data analysis. Because of lacking of some important contents, including relationships among experimental tasks, relationships among analytical methods, analytical results and their interpretations, *etc.*, these neuroimaging provenances cannot meet the requirements of systematic brain data management.

BI provenances have been proposed as BI-specific brain data provenances for realizing systematic brain data management. They are the metadata, which describe the origin and subsequent processing of various human brain data in the systematic BI study [16]. In our previous studies, a Data-Brain based approach has been developed to construct BI provenances [6]. The Data-Brain is a conceptual model of brain data, which represents functional relationships among

multiple human brain data sources, with respect to all major aspects and capabilities of human information processing system, for systematic investigation and understanding of human intelligence [16]. Owing to the BI-methodology-based modeling method, the Data-Brain and its own domain ontologies provide a knowledge base to guide the construction of BI provenances. By the Data-Brain-based approach, multi-aspect and multi-level data-related information can be integrated into BI provenances which connect the Data-Brain and heterogeneous brain data to form a brain data and knowledge base for meeting various requirements coming from the systematic BI study.

However, an important step in the Data-Brain-based development approach of BI provenances is to identify key concepts based on the Data-Brain, brain data and data-related information, for creating a conceptual framework of BI provenances. This means all of key concepts should be included in the Data-Brain before constructing BI provenances. Such a Data-Brain based approach often cannot be completed based on the existing prototype of the Data-Brain which only focuses on an induction-centric systematic BI study. The developers still need a provenance model which can provide a conceptual framework to tell the developers: what brain data related information should be obtained? How to organize the obtained information?

The Open Provenance Model (OPM) is a general provenance model to provide an effective conceptual framework for obtaining important information of biological logic origin and sequence processes [7,11]. By extending the OPM, a BI provenance model can be developed. The detail will be discussed in the following sections.

3 A Brain Informatics Provenance Model

As stated in our previous studies, BI provenances can be divided into data provenances and analysis provenances [16]. Data provenances describe the brain data origin and analysis provenances describe what processing on a brain dataset has been carried out.

The BI provenance model provides a conceptual framework for constructing data provenances and analysis provenances. It includes two types of model elements, basic elements and extended elements.

3.1 Basic Elements

During the systematic BI study, both experiments and data analysis consist of many human actions involved with actors, actions and results. Hence, based on the OPM, three basic elements of BI provenance model can be defined as follows.

Definition 1. *An* **Artifact**, *denoted by* **Ar**, *is an immutable piece of state used or produced during BI experiments or data analysis, which may have a physical embodiment in a physical object, such as a MRI scanning equipment "Siemens 3T Trio Tim Scanner", or a digital representation in a computer system, such*

as a neuroimaging analytical software "Statistical Parametric Mapping(SPM)". The artifact is represented by a circle, as shown in Figure 1.

Definition 2. *A* **Process**, *denoted by* **Pr**, *is an action or a series of actions performed on or caused by artifacts or others during BI experiments or data analysis. For example, an experiment is a process. The process is represented by a square, as shown in Figure 1.*

Definition 3. *An* **Agent**, *denoted by* **Ag**, *is a contextual entity acting as a catalyst of a process enabling, facilitating, controlling, or affecting its execution. For example, an experimental operator is an agent. The agent is represented by an octagon, as shown in Figure 1.*

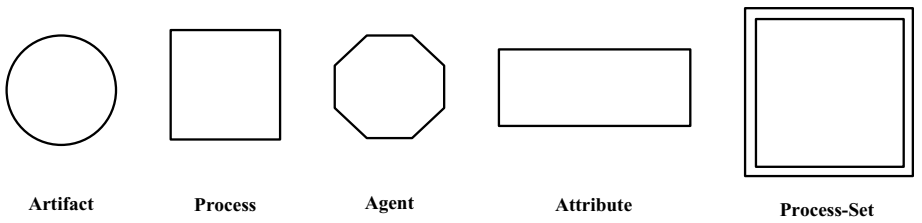

Fig. 1. The elements of the BI provenance model

3.2 Extended Elements

Three basic elements cannot meet the requirements of modeling BI data provenances and analysis provenances. Hence, two extended elements are defined in the BI provenance model.

Different artifacts, processes and agents have their own characteristics which are very important for identifying and understanding each type of artifacts, processes and agents. For describing these characteristics, an extended element Attribute is defined as follows.

Definition 4. *An* **Attribute**, *denoted by* **At**, *is a mapping*:

$$At : E \rightarrow C, S, T, N, \text{or } \varnothing$$

where E=$\{e \mid e$ is an Ar, Pr, or Ag$\}$, C is a set of characters, S is a set of strings, T is a set of texts, and N is a set of numbers, for describing a characteristic of artifacts, processes or agents. For example, the age is an Attribute which is a mapping between the set of the agents "operator" and the set of numbers. The Attribute is represented by a rectangle, as shown in Figure 1. $At(e)$ is the image of e under the mapping At and used to denote the value of attribute At of e.

There are many similar processes during systematic BI experiments and data analysis. For example, researchers often obtain brain data by a group of experiments which are same except for subjects. For describing such a similarity among processes, an extended element Process-Set is defined as follows.

Definition 5. *A* **Process-Set**, *denoted by* **PrS**, *is a set of processes*:

$$\{prs \mid \exists At, At(prs_i) = v \wedge prs_i \text{ is a } Pr, i = 1 \ldots n\},$$

where v is a character, string, text or number. For example, an experimental group is a Process-Set which is used to describe a group of experiments which are same except for subjects. The Process-Set is represented by two squares, as shown in Figure 1.

4 A Conceptual Framework of Brain Informatics Provenances

Data provenances describe the brain data origin by multi-aspect experiment information, including subject information, how experimental data were collected, and what instrument was used, *etc.* As shown in Figure 2, a general conceptual framework of data provenances can be described by using the BI provenance model. Table 1 gives major elements in this conceptual framework. All attributes are not included in this table because of limitation of space.

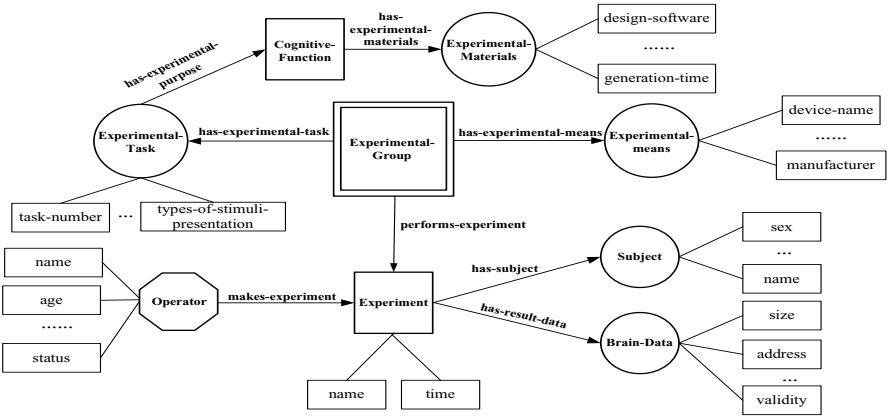

Fig. 2. A conceptual framework of data provenances

Figure 2 is only a general conceptual framework of data provenances. For describing a given dataset, it is necessary to construct a specific conceptual framework in which more specific artifacts, processes, agents, process-sets and attributes are used. The detail will be introduced by the case study in the next section.

Analysis provenances describe what processing in a brain dataset has been carried out, including what analytic tasks were performed, what experimental data were used, what data features were extracted, and so on. Figure 3

Table 1. All elements of a conceptual framework of data provenances

ID	TYPE	NAME	DESCRIPTION
001	Artifact	Experimental-Task	a group of tasks which need to be completed, such as the addition task "2+3=?"
002	Artifact	Subject	a man or woman who performs experimental tasks
003	Agent	Operator	a man or woman who carries out the experiments
004	Artifact	Experimental-Materials	a group of digital representations, such as figures, programs and texts, which are used to represent tasks
005	Process	Experiment	a virtual concept which is used to record the process information and integrate related concepts
006	Artifact	Brain-Data	experimental data which record physiological changes of brains during performing tasks
007	Artifact	Experimental-Means	a measuring device or technology which is used to collect brain data during the experimental process
008	Process	Cognitive-Function	a kind of capability of human brain which is used to complete experimental tasks
009	Process-Set	Experimental-Group	a group of experiments which are same except for subjects

represents a general conceptual framework of analysis provenances by using the BI provenance model. Major elements in this framework are introduced in Table 2.

Similar to data provenances, Figure 3 is only a general conceptual framework of analysis provenances. For describing a given data analysis, it is necessary to construct a specific conceptual framework for the corresponding analysis provenance.

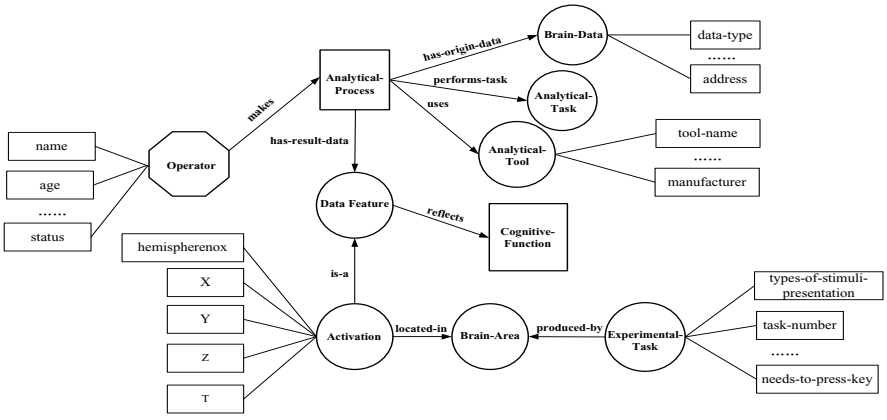

Fig. 3. A conceptual framework of analysis provenances

Table 2. All elements of a conceptual framework of analysis provenances

ID	TYPE	NAME	DESCRIPTION
001	Artifact	Experimental-Task	a group of tasks which need to be completed, such as the addition task "2+3=?"
002	Artifact	Brain-Data	experimental data which record physiological changes of subjects' brains
003	Process	Cognitive-Function	a kind of capability of human brain which is used to complete experimental tasks
004	Process	Analytical-Process	a virtual concept which is used to record the process information of BI data analysis and integrate analysis-related concepts
005	Artifact	Analytical-Task	a group of tasks which need to be completed during the analytical process
006	Artifact	Analytical-Tool	a software which is used to analyze brain data
007	Artifact	Data-Feature	a spatio-temporal characteristic of human information processing courses which is extracted from brain data
008	Artifact	Activation	a brain component or part which is reacted
009	Artifact	Brain-Area	a part in the brain
010	Agent	Operator	a man or woman who carries out data analysis

5 A Case Study in Thinking Centric Systematic Investigations

The BI study is data-driven and can be divided into four stages, question definition, experiment design, data analysis and result interpretation. In order to carry out the systematic BI methodology, the implementation of every stage should be based on a large number of experiences about experiments and data analysis. Before defining questions, researchers need to find similar studies and understand their experimental tasks, analytical methods and research results. Before designing experiments, researchers need to find similar experiments and understand their key experimental information, including types of experimental materials, the number of sessions, *etc.* Before analyzing data, researchers need to find similar analytical processes and understand their analytical information, including analytical methods, parameters, *etc.* Before interpreting results, researchers need to find related physiological characteristics of brain, including activated brain regions, functional connections, *etc.* However, it is difficult to complete above work only depending on individuals because of involving a large amount of knowledge about existing experiments and data analysis. BI provenances provide an effective way to support the above work. Their significance and usefulness will be introduced by the following case study.

Inductive reasoning is a kind of important human cognitive function. BI researchers have completed a series of induction studies, involved with 28 groups of experiments and 1130 subjects. The obtained data include fMRI(Functional

Magnetic Resonance Imaging) data, ERP(Event-Related Potential) data and eye-tracking data. A group of BI provenances were constructed for these data. For example, a data provenance was constructed for the fMRI dataset which was obtained by a group of experiments on numerical inductive reasoning [10]. Figure 4 is a fragment of the corresponding data provenance and describes the origin of a fMRI dataset which was obtained by one experiment in the experimental group. Zhao Hong is the operator of the experiment and a college student Li Pengyu is the subject. Two types of experimental tasks, including 30 induction tasks and 30 calculation tasks, were completed and experimental data were collected by the Siemens Trio Tim 3T. All BI provenances were represented by the RDF(Resource Description Framework) [13]. Based on these BI provenances and the induction-centric Data-Brain, the above four stages of the systematic BI study can be effectively supported by some SPARQL based queries [2,16]. For example, a SPARQL query Q1 shown in Figure 5 can be used to find similar experiments for understanding their experimental design during an induction-centric systematic BI study.

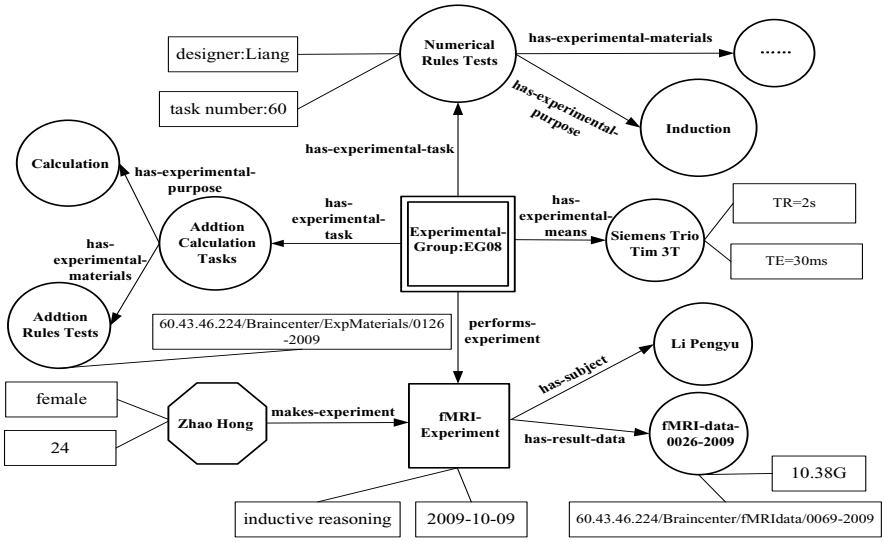

Fig. 4. The inductive reasoning construction of data provenances

In the Q1 "?Experimental_Task_URI waasb:has-experimental-purpose ?Cognitive_Function_URI. ?Cognitive_Function_URI rdf:type waasb:Reasoning." means that the similar experiments are the experiments whose experimental purposes are to study the cognitive function *Reasoning*, including its subclasses, such as *Induction* and *Deduction*. As shown in Table 3, though experimental purposes in data provenances were recorded as *Induction*, the corresponding experiments can still be found by reasoning based on data provenances and the

```
Q1: PREFIX waasb:
<http://www.semanticweb.org/ontologies/2011/11/DataBrain.owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT DISTINCT ?Experimental_Group ?Experimental_Task
?Cognitive_Function ?Types_Of_Stimuli_Presentation
?Equipment WHERE {
?Experimental_Group_URI waasb:name ?Experimental_Group.
?Experimental_Group_URI waasb:has-experimental-task ?Experimental_Task_URI.
?Experimental_Task_URI waasb:name ?Experimental_Task.
?Experimental_Task_URI
waasb:has-experimental-purpose ?Cognitive_Function_URI.
?Cognitive_Function_URI rdf:type waasb:Reasoning.
?Cognitive_Function_URI waasb:name ?Cognitive_Function.
?Experimental_Task_URI
waasb:types-of-stimuli-presentation ?Types_Of_Stimuli_Presentation.
?Experimental_Group_URI waasb:has-experimental-means ?Equipment_URI.
?Equipment_URI waasb:name ?Equipment.
}
ORDER BY ?Experimental_Group
```

Fig. 5. The SPARQL query Q1

Table 3. Results of the SPARQL query Q1

ID	Experimental _Group	Experimental_ Task	Cognitive _Function	Types_Of_Stimuli_ Presentation	Equipment
1	EG04	The reversed triangle inductive task (fMRI)	Induction	Synchronous	Siemens Trio Tim 3T
2	EG05	The reversed triangle inductive task (ERP)	Induction	Synchronous	Four 32 Channel BrainAmp MR Amplifiers
3	EG08	Numerical rules tests	Induction	Serial	Siemens Trio Tim 3T
4	EG11	Sentential inductive strength judgment	Induction	Serial	Siemens Trio Tim 3T
5	EG12	Sentential induction with multi-level preconditions	Induction	Serial	Siemens Trio Tim 3T
	...	...	...	...	...

Data-Brain because *Induction* is defined as a subclass of *Reasoning* in the Data-Brain. More complex rules can also be used to define the "similar", as stated in our previous studies [16].

6 Conclusions

BI provenances play an important role in the integration and synthetic utilization/mining of brain big data during the systematic BI study. This paper proposed a BI provenance model by extending the OPM. The case study in

thinking-centric systematic investigations shows usefulness of the proposed model for the systematic BI study. Furthermore, the obtained BI provenances can be used to support meta-analysis, provenances mining, the process planning of systematic brain data analysis, *etc.* All of these will be studied in our next work.

Acknowledgements. This work is supported by International Science & Technology Cooperation Program of China (2013DFA32180), National Key Basic Research Program of China (2014CB744605), National Natural Science Foundation of China (61272345), the CAS/SAFEA International Partnership Program for Creative Research Teams, and Open Foundation of Key Laboratory of Multimedia and Intelligent Software (Beijing University of Technology), Beijing.

References

1. Archer, W., Delcambre, L., Maier, D.: A Framework for Fine-grained Data Integration and Curation, with Provenance, in a Dataspace. In: Proceedings of the First Workshop on Theory and Practice of Provenance, pp. 1–10 (2009)
2. Barbieri, D.F., Braga, D., Ceri, S., Della Valle, E., Grossniklaus, M.: Querying RDF Streams with C-SPARQL. Special Interest Group on Management of Data Record 39(1), 20–26 (2010)
3. Cheney, J., Chiticariu, L., Tan, W.C.: Provenance in Databases: Why, How and Where. Foundations and Trends in Databases 1(4), 379–474 (2007)
4. Cui, Y.W., Widom, J., Wiener, J.L.: Tracing the Lineage of View Data in a Warehousing Environment. ACM Transactions on Database Systems 25(2), 179–227 (2000)
5. Chen, J.H., Zhong, N.: Data-Brain Modeling Based on Brain Informatics Methodology. In: Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2008), Sydney, NSW, Australia, pp. 41–47. IEEE Computer Society (2008)
6. Chen, J.H., Zhong, N., Liang, P.P.: Data-Brain Driven Systematic Human Brain Data Analysis: A Case Study in Numerical Inductive Reasoning Centric Investigation. Cognitive Systems Research 15(1), 17–32 (2011)
7. Dai, C., Lim, H.S., Bertino, E., Moon, Y.S.: Assessing the Trustworthiness of Location Data Based on Provenance. In: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems, pp. 276–285 (2009)
8. Foster, J., Green, J., Tannen, V.: Annotated XML: Queries and Provenance. In: Proceedings of the 27th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, pp. 271–280 (2008)
9. Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., Hill, D.P., Kania, R., Schaeffer, M., St Pierre, S., Twigger, S., White, O., Rhee, S.Y.: Big Data: The Future of Biocuration. Nature 455(7209), 47–50 (2008)
10. Lu, S.F., Liang, P.P., Yang, Y.H., Li, K.C.: Recruitment of the Pre-motor Area in Human Inductive Reasoning: an fMRI Study. Cognitive Systems Research 11(1), 74–80 (2010)
11. Moreau, L., Freire, J., Futrelle, J., McGrath, R.E., Myers, J., Paulson, P.: The Open Provenance Model. Technical Report 14979, School of Electronics and Computer Science, University of Southampton, pp. 1–8 (2007)

12. MacKenzie-Graham, A.J., Van Horn, J.D., Woods, R.P., Crawford, K.L., Toga, A.W.: Provenance in Neuroimaging. NeuroImage 42(1), 178–195 (2008)
13. Ni, W., Chong, Z.H., Shu, H., Bao, J.J., Zhou, A.Y.: Evaluation of RDF Queries via Equivalence. Frontiers of Computer Science 7(1), 20–33 (2013)
14. Zhong, N., Liu, J.M., Yao, Y.Y.: In Search of the Wisdom Web. Special Issue on Web Intelligence (WI), IEEE Computer 35(11), 27–31 (2002)
15. Zhong, N., Bradshaw, J.M., Liu, J., Taylor, J.G.: Brain Informatics. Special Issue on Brain Informatics, IEEE Intelligent Systems 26(5), 16–21 (2011)
16. Zhong, N., Chen, J.H.: Constructing a New-style Conceptual Model of Brain Data for Systematic Brain Informatics. IEEE Transactions on Knowledge and Data Engineering 24(12), 2127–2142 (2011)