Organizing Knowledge for Web Retrieval using SKOS: A Case Study in Human Protein Chain

Abhijit Dasgupta
Assistant Librarian
Indian Association for the Cultivation of Science (IACS)
(Under Dept. of Science and Technology, GOI)
2A & B, Raja S C Mullick Road, Jadavpur, Kolkata, India
abhijit_60@hotmail.com; library@iacs.res.in

D S Rath
Reader
Dept. of Library & Information Science
Vidyasagar University
Midnapore, West Bengal, India
durga_s_rath@rocketmail.com

Abstract
Effective knowledge management is the most challenging task today to organize and control the millions of web resources in any scholarly publications. An effort is made to map human protein chain against different neurological disorders. After analyzing the facets in this domain, a thesaurus is constructed, relational structure of SKOS is made and finally converted into XML:RDF compliant format for knowledge representation, manipulation, interoperability and effective retrieval.

Keywords: *SKOS (Simple Knowledge Organization System), RDF (Resource Description Framework), Ontology, Semantic mapping, Knowledge Management*

Introduction
The knowledge driven society has not only changed the very character of information usage pattern where the value addition has become a pertinent criteria of any information retrieval product, the management of so called knowledge resources has also become extremely complex due to interdisciplinary character of the universe of subjects. Considering the growth pattern of any thrust area of research topic and its multi-subject approach pattern, the management of published data/information on that very topic faces a daunting task for any people or organization.

To overcome the problem of relevance of the retrieved information from the internet, the semantic web technology has come into the focal point of study, where efforts are being made to arrange the metadata (hyperlinked index of any digital textual/graphic matter) into an ontological relational structure or we can say in a classified structure where facets are in a relational mapping with other facets to some extent similar as that of a thesaurus structure. So, when a scholarly approach will be taken towards search and retrieval of any literature from the web, the search against any particular or group of metadata will be restricted within the relevant domain of subject databases. Basically semantic is the study of meaning. The semantic has come from the Greek word *semantikos*, which means "significant meaning"; semantic web technologies initiate separate meaning from data, document content and application code. Semantic web basically a technology based on open standards. Semantic technologies represent meaning based on ontology and provide reasoning through the relational structure, rules, logic and conditions represented in those ontologies.(Balani, 2009)

SKOS (Simple Knowledge Organization System)
The SKOS core mapping concept had been framed and developed by the W3C (World Wide Web consortia). It has been already discussed that the basic objective of the SKOS mapping is to fulfill the concept of semantic web technology, where the metadata of any web literature will be arranged in an ontological as well as semantic relationship. We can infer that the semantic web is a mesh of concept linked up in such a way as to be easily processible by machine, on a global scale (Dumbill, 2000). Though the semantic web technologies are still in their infancies or in the early stage of adolescence, but specific researches are going on the areas of ontology, metadata structures, vocabularies, and resource description framework (RDF). SKOS is an area of work concentrating on developing specifications and standards to support the use of knowledge organization systems (KOS), such as thesauri, classification schemes, subject heading systems and taxonomies within the framework of semantic web. (Miles, A. and others , 2008)

Semantic Web Technologies
World Wide Web Consortium (W3C) has designed a set of web languages to express the meaning of the information resources on the web and each of these languages is the extension of the former:

- XML (Extensible Mark-up Language)
- XML Schema
- RDF (Resource Descriptive Framework)
- RDF Schema
- OWL (Web Ontology Language)

Dependence and Extension Model of the Semantic Web Languages

Extensible Mark-up Language (XML)/Schema --- → extended by --- → Resource Descriptive Framework (RDF)/Schema --- → extended by --- → Web Ontology Language (OWL)

Idea behind the subject for semantic mapping of Human Protein Chain

This is a revolutionary development in the field of molecular biology and bioinformatics, where nearly 30,000 different proteins of human cell those are being identified and mapped. The immense potential of the method in clinical diagnosis was described by E. Jellum (University of Oslo) and L. Anderson. Clearly a catalogue of human gene products could make it possible to detect metabolic irregularities and diseases, such as cancer, which affect gene expression.

Protein interaction maps are powerful tools for determining the cellular functions of genes. Till date large scale protein interaction maps have been generated for several invertebrates species, but similar scale has not been yet described for any mammals, as because several physical interactions are conserved between species, it should be possible to infer information about human protein interactions (and protein function) using model organism protein interaction datasets (Schulz, G. E, 1979)

Knowledge Organization of the web resources dealing with Human Protein Chain

The human protein chain sequence itself is an entity of embedded large scale of knowledge resources in relation to manifestation of several major diseases. So, the human protein chains of the different organs of the brain have been dealt as the metadata, where each metadata (i.e., protein chain sequence) has been arranged in an ontological semantic order. Thus, in an actual web retrieval scenario this protein chain metadata will plug-in the related knowledge resources in a digital library platform enhancing the quality of relevancy. Here the aspects of knowledge organization and management in relation to human protein chain map have been dealt in three areas:

i) Storage of the knowledge resources on human protein chain;

ii) Maneuvering of the knowledge resources on the said subject;

iii) Effective search and retrieval of the knowledge resources through SKOS ontological mapping and XML: RDF graph against the human protein chain.

Here the protein chain sequence codes, assigned by the molecular biologists, are itself a so called globally accepted classified structure of protein chain. So, each protein chain code along with its abbreviated form of diseases like P1236045Alz can be dealt as a standard metadata (Angelis & others , 2004)

Semantic Relational Structure of the Human Tissue Protein against particular Neurological

Disease/Disorder (Alzheimer's Disease and Parkinson's Disease) against particular Lobe of theNeurological Organ/Tissue

(Note: First order – Name of the disease --- → Second order – Location organ/tissue of the disease --- →

Third order- Name of the Human Protein Chain code against that particular organ/tissue) (*http:// www.biochain.com/biochain*)

BT Neurological Disease/Disorder - Human

NT Alzheimer's Disease (Alz)
NT Alz Location Tissue – Brain
RT Protein Chain Code – P1236035Alz
NT Alz Location Tissue – Brain – Amygdala
RT Protein Chain Code – P1236036Alz
NT Alz Location Tissue – Brain – Cerebellum
RT Protein Chain Code – P1236039Alz
NT Alz Location Tissue – Brain – Cor/ Call
RT Protein Chain Code – P1236045Alz
NT Alz Location Tissue – Brain – Frontal Lobe
RT Protein Chain Code – P1236051Alz
NT Alz Location Tissue – Brain – Hippocampus
RT Protein Chain Code – P1236052Alz
NT Alz Location Tissue – Brain – Med Oblongata
RT Protein Chain Code – P1236057Alz
NT Alz Location Tissue – Brain – Occipital Lobe
RT Protein Chain Code – P1236062Alz
NT Alz Location Tissue – Brain - Parietal Lobe
RT Protein Chain Code – P1236066Alz
NT Alz Location Tissue – Brain – Pons
RT Protein Chain Code – P1236071Alz
NT Alz Location Tissue – Brain – Postcentral Gyrus
RT Protein Chain Code – P1236072Alz
NT Alz Location Tissue – Brain – Precentral Gyrus
RT Protein Chain Code – P1236073Alz

NT Alzheimer's Disease (Alz) contd...
NT Alz Location Tissue – Brain – Temporal Lobe
RT Protein Chain Code – P1236078Alz
NT Alz Location Tissue – Brain – Thalamus
RT Protein Chain Code – P1236079Alz

NT Parkinson's Disease (Par)
NT Par Location Tissue – Brain
RT Protein Chain Code – P1236035Par
NT Par Location Tissue – Brain – Amygdala
RT Protein Chain Code – P1236036Par

NT Par Location Tissue – Brain – Cerebellum
RT Protein Chain Code – P1236039Par
NT Par Location Tissue – Brain – Corpus Callosum
RT Protein Chain Code – P1236045Par
NT Par Location Tissue – Brain – Frontal Lobe
RT Protein Chain Code – P1236051Par
NT Par Location Tissue – Brain – Medulla Oblongata
RT Protein Chain Code – P1236057Par
NT Par Location Tissue – Brain – Occipital Tissue
RT Protein Chain Code – P1236062Par
NT Par Location Tissue – Brain – Parietal Lobe
RT Protein Chain Code – P1236066Par
NT Par Location Tissue – Brain – Pons
RT Protein Chain Code – P1236071Par
NT Par Location Tissue – Brain – Postcentral Gyrus
RT Protein Chain Code – P1236072Par
NT Par Location Tissue – Brain – Precentral Gyrus
RT Protein Chain Code – P1236073Par
NT Par Location Tissue – Brain – Temporal Lobe
RT Protein Chain Code – P1236078Par
NT Par Location Tissue – Brain – Thalamus
RT Protein Chain Code – P1236079Par

Transferring the facets/isolates of Human Protein Chain against respective Neurological diseases into SKOS Core Semantic Mapping – RDF Graph (Standard followed – SKOS Core Guide: *http://www.w3.org/TR/2005/WD-swbp-skos-core-guide-20051102*)

SKOS core provides a model for expressing the basic structure and content of the concept schemes such as thesauri, classification schemes, subject heading lists, taxonomies, and other types of controlled vocabularies, and also the concept schemes embedded in glossaries and terminologies. SKOS core is an application of the Resource Description Framework (RDF). RDF provides simple data formalism for talking about things, their properties, inter-relationships, and categories (Classes).RDF semantics for its formal mathematical basis, and RDF syntax for details of the RDF/XML documents format used to exchange RDF data. The SKOS core vocabulary is a set of RDF properties and RDFS classes that can be used to express the content and structure of a concept scheme as a RDF graph. (Mount, D.W., 2001)

Efforts have been taken to transfer the respective protein chain sequence against the respective neurological disorder/disease into SKOS core semantic mapping – RDF graph.

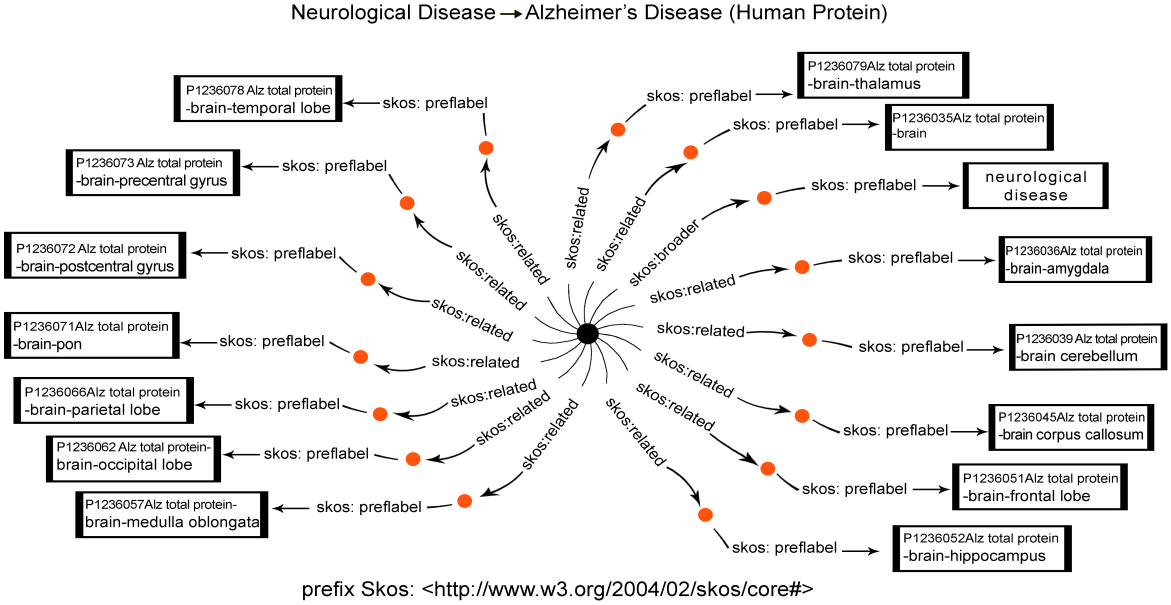
Transformation of the SKOS Core Semantic Mapping RDF graph of Human Protein Chain responsible for Neurological Disorders/ Diseases into XML: RDF Compliant Program

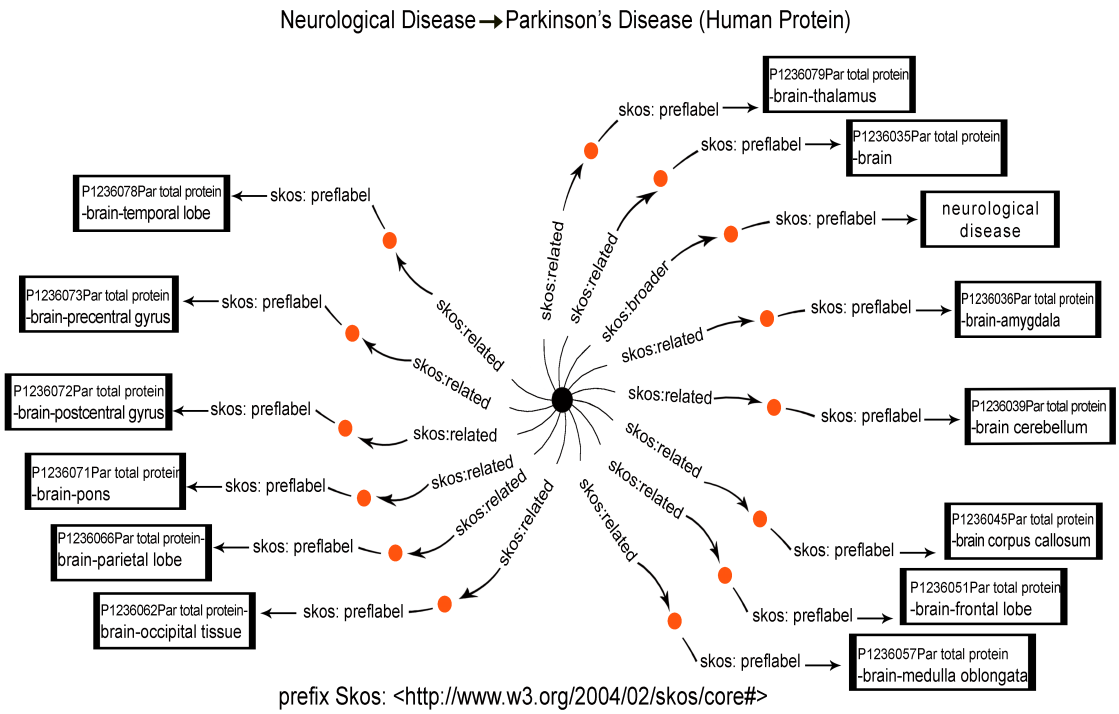
Protein Chain of a single Major Neurological Diseases have been taken for XML RDF Compliant Program – i.e. Alzheimer's Disease and Parkinson's Disease Standard Followed: SKOS Core Guide (W3C Working Draft – Nov., 2005)URL: http://www.w3.org/TR/2005/WD-swbp-skos-core-guide-20051102

(Note: The original program file written in Notepad given in Annexure – I/ As it is just a pasted copy in doc file of the notepad XML: RDF program, some distortion of the indention have been occurred, originally the program should be opened in the Notepad file)

XML:RDF compliant program file copy of the Alzheimer's disease – Human Brain - Total Protein (Prototype Model)

<?xml version="1.0" encoding="UTF-8"?>






```
<rdf:RDF
    xmlns:rdf="http://www.w3c.org/1999/02/22-rdf-syntax-ns#"
   xmlns:skos="http://www.w3c.org/2004/02/skos/core#">
<skos:concept rdf:about="http://www.example.com/concepts#alzheimer's disease - human brain - total protein">
    <skos:preflabel>alzheimer's disease - human brain - total protein</skos:preflabel>
    <skos:broader rdf:resource="http://www.example.com/concepts#neurological disease - human - total protein"/>
</skos:concept>
<skos:concept rdf:about="http://www.example.com/concepts#P1236035Alz - brain">
    <skos:preflabel>P1236035Alz - brain</skos:preflabel>
    <skos:related rdf:resource="http://www.exmaple.com/concepts#alzheimer's disease - human brain - total protein"/>
</skos:concept>
<skos:concept rdf:about="http://www.example.com/concepts#P1236036Alz - brain - amygdala">
    <skos:preflabel>P1236036Alz - brain - amygdala</skos:preflabel>
    <skos:related rdf:resource="http://www.example.com/concepts#alzheimer's disease - human brain - total protein"/>
</skos:concept>
<skos:concept rdf:about="http://www.example.com/concepts#P1236039Alz - brain - cerebellum">
    <skos:preflabel>P1236039Alz - brain - cerebellum</skos:preflabel>
    <skos:related rdf:resource="http://www.example.com/concepts#alzheimer's disease - human brain - total protein"/>
</skos:concept>
<skos:concept rdf:about="http://www.example.com/concepts#P1236045Alz - brain - corpus callosum">
    <skos:preflabel>P1236045Alz - brain - corpus callosum</skos:preflabel>
    <skos:related rdf:resource="http://www.example.com/concepts#alzheimer's disease - human brain - total protein"/>
</skos:concept>
<skos:concept rdf:about="http://www.example.com/concepts#P1236051Alz - brain - frontal lobe">
    <skos:preflabel>P1236051Alz - brain - frontal lobe</skos:preflabel>
    <skos:related rdf:resource="http://www.example.com/concepts#alzheimer's disease - human brain - total protein"/>
</skos:concept>
<skos:concept rdf:about="http://www.example.com/concepts#P1236052Alz - brain - hippocampus">
    <skos:preflabel>P1236052Alz - brain - hippocampus</skos:preflabel>
    <skos:related rdf:resource="http://www.example.com/concepts#alzheimer's disease - human brain - total protein"/>
</skos:concept>
<skos:concept rdf:about="http://www.example.com/concepts#P1236057Alz - brain - medulla oblongata">
    <skos:preflabel>P1236057Alz - brain - medulla oblongata</skos:preflabel>
    <skos:related rdf:resource="http://www.example.com/concepts#alzheimer's disease - human brain - total protein"/>
</skos:concept>
<skos:concept rdf:about="http://www.example.com/concepts#P1236062Alz - brain - occipital lobe">
    <skos:preflabel>P1236062Alz - brain - occipital lobe</skos:preflabel>
```

<skos:related rdf:resource="http://www.example.com/concepts#alzheimer's disease - human brain - total protein"/>
</skos:concept>
<skos:concept rdf:about="http://www.exmaple.com/concepts#P1236066Alz - brain - parietal lobe">
<skos:preflabel>P1236066Alz - brain - parietal lobe</skos:preflabel>
<skos:related rdf:resource="http://www.example.com/concepts#alzheimer's disease - human brain - total protein"/>
</skos:concept>
<skos:concept rdf:about="http://www.example.com/concepts#P1236071Alz - brain - pons">
<skos:preflabel>P1236071Alz - brain - pons</skos:preflabel>
<skos:related rdf:resource="http://www.example.com/concepts#alzheimer's disease - human brain - total protein"/>
</skos:concept>
<skos:concept rdf:about="http://www.example.com/concepts#P1236072Alz - brain - postcentral gyrus">
<skos:preflabel>P1236072Alz - brain - postcentral gyrus</skos:preflabel>
<skos:related rdf:resource="http://www.example.com/concepts#alzheimer's disease - human brain - total protein"/>
</skos:concept>
<skos:concept rdf:about="http://www.example.com/concepts#P1236073Alz - brain - precentral gyrus">
<skos:preflabel>P1236073Alz - brain - precentral gyrus</skos:preflabel>
<skos:related rdf:resource="http://www.example.com/concepts#alzheimer's disease - human brain - total protein"/>
</skos:concept>
<skos:concept rdf:about="http://www.example.com/concepts#P1236078Alz - brain - temporal lobe">
<skos:preflabel>P1236078Alz - brain - temporal lobe</skos:preflabel>
<skos:related rdf:resource="http://www.example.com/concepts#alzheimer's disease - human brain - total protein"/>
</skos:concept>
<skos:concept rdf:about="http://www.example.com/concepts#P1236079Alz - brain - thalamus">
<skos:preflabel>P1236079Alz - brain - thalamus</skos:preflabel>
<skos:related rdf:resource="http://www.example.com/concepts#alzheimer's disease - human brain - total protein"/>
</skos:concept>
</rdf:RDF>

Conclusion

Knowledge organization of the web resources will be a much more complex and challenging task in the near future owing to uncontrolled growth of the electronic resources and to maintain proper relevancy of the retrieved web literatures. Semantic web is meant to bring order in the chaotic disorderliness of information organization or we can say knowledge organization. Unless context specific information are available, information in this digital era appears as if searching needle in the hay stacks. Interdisciplinary research made the problem of relevance more unmanageable. In preparation of thesaurus, question arises on standardization of terms, universal hierarchy in arrangement, contextual differences, etc. We will take resort to domain experts to sort out these issues, besides the literary warrant of the domain. However, once the researcher finds the protein chain directly, those are responsible for any neurological disease/disorder, he/she can concentrate on the sequence within the chain, that reduce the time gap in search process, and obviously most pertinent to the problem in hand. Moreover as this thrust area of research on human protein chain against any particular disease has already ushered a new horizon in the field of drug design and bioinformatics, it has strongly felt that to manage the robust digital micro-level knowledge resources on this area in near future, the semantic technology application is absolutely necessary.

References

1. Angelis, G. and others : Mechanisms for controlling access in the global grid environment. (*Internet Research, v. 14, pp.347-352, 2004*)
2. Attwood, T. (and) Parry-Smith, D.: Introduction to Bioinformatics. 1999.
3. Balani, Naveen: The Future of the Web is Semantic – Ontologies form the backbone of a whole new way to understand online data (URL: *http://www.ibm.com/developerworks/web/library/wa-semweb/*) (Retrieved: 2/07/2009)
4. Biochain – Genomic DNA, cDNA, RNA, Protein… (URL: *http://www.biochain.com/biochain*) (Retrieved: 17/06/2009)
5. Bioinformatics Thesaurus: Bioinformatics Resource Portal (URL: *http://www.geocities.com/bioinformatics web/thesurus.html*) (Retrieved: 24/06/2009)
6. Clark, Brian, F. C: Towards a Total Human Protein Map (*Nature, v.292, pp.491-492, 1981*)
7. Klyne, G. & Carroll, J. Resource Descriptive Framework (RDF): Concepts and Abstract Syntax. W3C. 2004 (*http://www.w3.org/TR/rdf-concepts/*) (Retrieved: 9/08/2008)
8. Miles, A. and others. SKOS Core Guide, 2nd W3C Public Working Draft. W3C. 2004 (*http://www.w3.org/2001/sw/Europe/reports/thes/1.0/guide/*) (Retrieved: 24/07/2008)
9. Mount, D. W.: Bioinformatics: Sequence and Genome Analysis. 2001
10. Rowley, J. E. and Hartley, R.: Organizing Knowledge: an introduction to managing access to information. 2008
11. Rual, J-F., and others: Towards a Proteome Scale map of the Human Protein-Protein Interaction Network (*Nature, v.437, pp.1173-1178, 2005*)
12. SKOS Core Guide: W3C Working Draft, Nov., 2005 (*http://www.w3.org/TR/2005/WD-swbp-skos-core-guide-20051102*) (Retrieved: 11/07/2008)
13. Schulz, G. E.: Principles of Protein Structure. 1979.