ASSOCIAÇÃO BRASILEIRA DE ENERGIA NUCLEAR - ABEN
2019 International Nuclear Atlantic Conference - INAC 2019
Santos, SP, Brazil, October 21-25, 2019

Agência Brasileira do ISBN
ISBN 978-85-99141-08-3
9 788599 141083

DEVELOPMENT OF A SEMANTIC LAYER FOR A DATA REPOSITORY PROTOTYPE FOR NEUTRON ACTIVATION ANALYSIS DOMAIN DATA

Glauber Mauch de Carvalho, Renato Semmler,
and Mário Olímpio de Menezes

Nuclear and Energy Research Institute (IPEN/CNEN)
Research Reactor Center (CERPq)
Av. Professor Lineu Prestes 2242
05508-000 São Paulo, SP
glauber_mauch@usp.br,
rsemmler@ipen.br
momenezes@usp.br

ABSTRACT

In order to provide greater transparency for scientific research and the results achieved, a great effort is ongoing to make available scientific data repositories, which allow for different researchers to validate, reproduce and reuse third party scientific data. With the always increasing use of technology in all kinds of scientific facilities, a growing amount of data is collected even from simple experiments. This scenario presents a new paradigm: the understanding of third party data easily found on data repositories. Consequently, being able to do useful searches on these data repositories poses a new challenge for traditional search engines. In this work, the approach taken to help solve these problems was to propose a semantic layer for scientific data repository. By using ontologies, with appropriated integration of traditional search mechanisms, it will be easier for users to find related data that could be used in their work, improving the overall scientific yield. In order to achieve this goal, a ontology was developed, using the Protègè software, for the Neutron Activation Analysis (NAA) data domain. This ontology was validated by experts from NAA Laboratory of the Reactor Research Center (CERPq) at the Nuclear and Energy Research Institute (IPEN-CNEN/SP). A prototype of a semantic data repository is, thus, being developed using the Django web development framework. RDFlib, a software library written in Python is being used to allow the integration of semantic operations, based on the NAA ontology, with the relational database layer provided by Django.

1. INTRODUCTION

Science can be understood as any competence acquired through study, research or practical development supported by principles based on reflection, observation and experimentation, culminating in theories that make it possible to be elaborated, refined, or renounced, so that quantity and the quality of information is preserved.

One of the foundations of science is the possibility of the reproduction of scientific research results by independent researchers, making it possible to validate methods, results and their conclusions. Due to the nowadays unmeasured proportion of scientific data production ("big data"), it is necessary to implement systematic methods of storage, curation and availabilization of data; in this context, a new scientific paradigm emerges:

e-Science, called “The Fourth Paradigm of Scientific Exploration,” which distinguishes data-intensive science from traditional computational science [1].

E-Science is related to the discovery and sharing of knowledge in the form of experimental data, rich theoretical vocabularies, reusable publications and services that are viable to the scientific community [2]. The complexity and abundance of data resources in an e-Science environment requires support for knowledge and metadata management since it’s a known fact that data commonly is difficult to share, find, access, interpret, or reuse [2].

The availability of data underlying published scientific articles has been established as a requirement by some innovative journals in the scientific field [2] as well as by a growing number of science funding agencies. Published data must be accompanied by knowledge about the data (metadata), including information about the data production, the methods, algorithms or other techniques employed to obtain those data as well as the all the analyzes that culminated in the publication to which they are associated.

In order to meet the demand of this new paradigm of science, mathematical models, digital repositories and data management, new hardware, software, protocols, tools and services, as well as several other initiatives have been applied to improve the infrastructure that supports sharing, findability, accessibility and reuse of research data. All this effort led to representatives of academia, industry, funding agencies, and publishers or publishers work together to design and endorse a concise and measurable set of principles, called FAIR (findability, accessibility, interoperability and reuse) Data Principles, a set of guides to improve the reuse of research data [3].

One of the main tools used to improve the sharing of data is by means of Data Repositories. However, data repositories by them selves may have limited reach due to lack of data format standards as well as lack of standardized data access protocols, which may result in problems for data discovery and reuse for both, humans and computer systems.

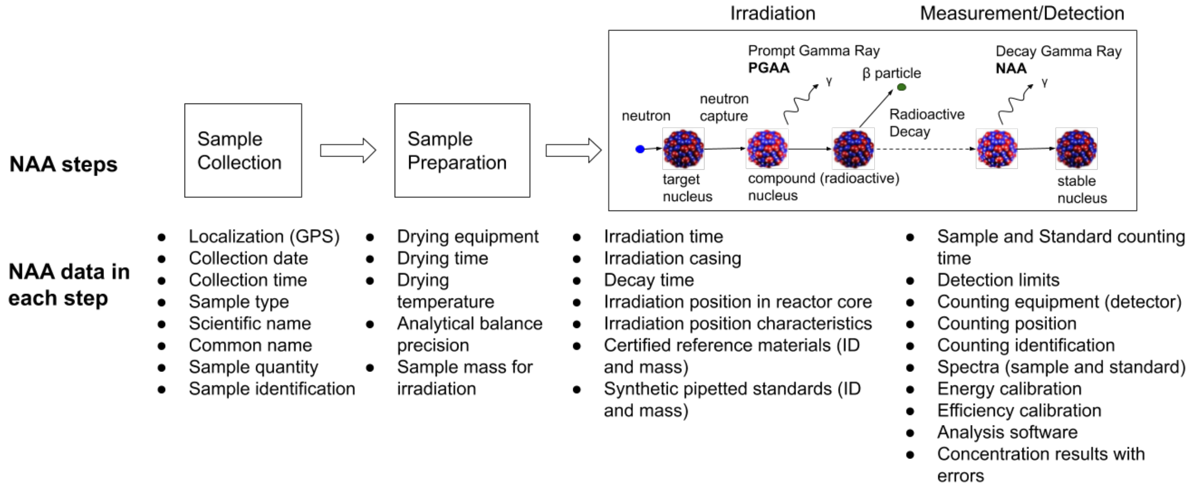
Because of these problems in traditional repositories, semantic technologies are increasingly gaining ground in e-Science. Methodologies, tools and other semantic-based components (ontologies) are increasingly being used. Knowledge modeling, logic-based hypothesis verification, semantic data integration, application compositions and knowledge discovery, and integrated data analysis for different knowledge domains are increasingly present in these contexts [4].

1.1. Neutron Activation Analysis - NAA

The principle of neutron activation analysis is the interaction of a given material with neutrons, which induces a nuclear reaction in an atom of a target element. The reaction product is detected and can be quantified by measuring its radioactive emission, that might include a prompt gamma ray, decay gamma ray as well as emission of a particle, that is, by its decay properties. For NAA, several nuclear reactions are possible depending on the target nucleus and the neutron energy [5].

In the CERPq Neutron Activation Laboratory, the commonly used NAA method is the comparative method, in which a sample with known mass is irradiated simultaneously with a standard, with known mass and element concentrations, and after a certain period of time, the intensity and energy of both standard and sample gamma ray spectra are measured. Comparison between specific activities induced in the standard and in the sample is the basis for calculating the element concentrations in the sample.

Each experiment of Neutron Activation Analysis (NAA) generates a large amount of data, some of which are shown in Figure 1.



- Localization (GPS)
- Collection date
- Collection time
- Sample type
- Scientific name
- Common name
- Sample quantity
- Sample identification

- Drying equipment
- Drying time
- Drying temperature
- Analytical balance precision
- Sample mass for irradiation

- Irradiation time
- Irradiation casing
- Decay time
- Irradiation position in reactor core
- Irradiation position characteristics
- Certified reference materials (ID and mass)
- Synthetic pipetted standards (ID and mass)

- Sample and Standard counting time
- Detection limits
- Counting equipment (detector)
- Counting position
- Counting identification
- Spectra (sample and standard)
- Energy calibration
- Efficiency calibration
- Analysis software
- Concentration results with errors

Figure 1: Data generated in each step of the NAA

Much of these data are not properly stored, curated and made available to other researchers. This scenario, which might be common to other IPEN Laboratories, has resulted in the initiative of the e-Science project at IPEN-CNEN/SP, and the creation of a institutional data repository.

1.2. Ontology Construction

Ontology construction demands great effort, discipline and a rigorous approach. All activities, systemic methodologies and processes involved compose the main subject of the Ontology Engineering. The main methodologies used to build ontologies can be found in the literature [6, 7, 8].

In this study, a hybrid methodology was employed, using tools like the Extended Language Lexicon [9], some concepts derived from Ontology Development 101[7] as well as from Rautenberg [8]. This resulted in a three steps process, described following:

1. Construction of the Extended Language Lexicon (ELL)

   "The Extended Language Lexicon (ELL) is a simple representation language. It is made up of only three basic entities: Term, Notion, and Impact; Terms might be further classified into "Subject", "Object", "State" and "Verb". ELL is a Requirements Engineering (a software engineering discipline) representation language that aims to map the vocabulary used in the Information Universe (IU), that is, the context in which the software should be developed and operated. The Information Universe includes all sources of information and all people related to the software. Table 1 presents the rules used to model the ELL.

Table 1: ELL formation rules

Term type	Notion	Impact
Subject	Who is the subject?	Which actions are performed?
Verb	Who does, when it happens and which procedures are involved	Which are the impacts of the action into the environment (IU), that is, what are other actions happening also, and which are the resulting states.
Object	Define the object and identify other related objects	Actions applied to the object
State	What it means and which actions are responsible for this state	Identify other possible states and actions that can follow from the current described state.

2. Lexical-Ontology Mapping ELL Analysis

   This step aims to map the terms of ELL into the elements of ontology: "object" and "subject" type terms are mapped into concepts, also called classes; "verb" type terms are mapped to properties and "state"-like type terms are mapped to concepts or properties; the notion of each term is mapped to the description of its concept; impact (description) is scanned to search for other terms already present or to be inserted into the ELL.

3. Class Hierarchy Construction

   This step consists of the analysis of the mapping done in the previous step, identifying all classes, and arranging them hierarchically. The essential difference between the Dictionaries, Controlled Vocabulary, Thesaurus or Lexical representations and Ontologies is the structure in which information is organized. In dictionaries, controlled vocabularies, thesaurus or lexical representations, the information is contained in a plane, while in Ontologies it is arranged in hierarchies. The classes of an ontology are structurally related through the specialization relationship. At the top of ontology is the most generic term, and at each level down, a more specific term.

There are several tools for editing ontologies such as: SWOOP, OntoStudio, NeOn Toolkit, TopBraid Composer and Protégé, which is one of the most popular tools. Protégé allows one to build the ontology graphically and save it in different formats like XML / RDF, Tortoise, RDF and RDFS (RDF Schema), Web Ontology Language (OWL) and eXtensible Markup Language (XML). Protégé is maintained by a strong community of collaborators, universities and governments where it's used to develop solutions in the most diverse areas of knowledge engineering [10].

In this study, we present the first results of a ongoing project regarding the construction of a semantic repository prototype for Neutron Activation Analysis (NAA) research data from the Neutron Activation Laboratory at the Research Reactor Center (CERPq), IPEN-CNEN/SP. By using a semantic layer in a data repository, we hope to enhance and increase the reuse of shared data, making it not only possible to validate methods, results and conclusions, but also to advance other researches that can benefit from such data.

2. MATERIALS AND METHODS

2.1. Construction of the Extended Language Lexicon (ELL)

As seen in Table 1, the Extended Language Lexicon (ELL) is composed of three basic entities: Term, Notion and Impact. Terms might be further classified into "Subject", "Object", "State" and "Verb". In this step we aim to map the vocabulary used in the Neutron Activation Analysis "Information Universe" into the ELL. This universe must include all sources of information and all persons related to the domain. The construction of the ELL is performed the two steps: Information Survey and Modeling.

In order to collect the pertinent terms within the domain of the NAA area, an extensive bibliographical survey was performed and several thesis and dissertations published by authors of the NAA Laboratory were selected; they were carefully read to find all terms that are related to data produced in each specific research. These terms were organized by source and by stage in the NAA process. Then, some interviews with specialists of the area were conducted to check if those terms were pertinents, if some terms were missing, etc. One of the findings in this step was the non uniform nomenclature used by different authors for the same concept, even belonging to the same laboratory, something that could confuse beginner students; all those terms with some ambiguity were solved after a few interactions. For the modeling of the ELL, the terms collected were organized in a spreadsheet to facilitate its manipulation; the model layout, with some of the terms are shown in Table 2

2.2. Construction of the NAA Ontology

After defining the terms, the ontology was modeled through the Protégé editor, and saved in the OWL (Ontology Web Language) format. Some metrics of the resulting ontology are: 43 classes, 31 object properties, 97 data properties and 127 subclasses.

Figures 2 and 3 present a small portion of the Neutron Activation Analysis ontology, respectively, some classes related to the Standards used in the comparative method of NAA (Certified Reference Material, Pipetted Standard) and some classes related to the Samples (Aliquot, Irradiated Aliquot, etc). In these figures the relationships, as exported by the Protégé software, are colored according to:

- **blue lines** — "type of" or "subclass", that is, a specialization relationship.
- **orange/yellow/violet/green dashed lines** — "property" relationship.

Table 2: Term, Notion and Impact of NAA domain

Term	Type	Notion	Impact (description)
Sample Identification	Subject	Identify the collected sample	Unique identification for the collected sample
Location	State	GPS coordinates of sample collection	Precisely indicate the location of sample collection
Sample mass	Object	Sample mass for irradiation	Final sample mass for irradiation
Certified reference material (CRM)	Object	Concentration of each element in the CRM	Knwon concentrations used to calculate the sample unknown element concentrations
Pipetted standard	Object	Concentration of elements in the Pipetted standard	Known concentrations used to calculate the sample unkown element concentrations
Irradiation time	State	Irradiation time	How long sample were irradiated
Decay time	State	Decay time	Duration of the radioactive decay of sample and standard after irradiated
Element concentrations	Subject	Quantities of elements in the sample	Quantity of each element, as determined by the NAA, in the sample

And the **arrow** points toward the specialization class in the hierarchy, so the reading is done backwards, for instance, in Figure 2, “Reference Material” is a **type of** “Non Irradiated Standard”. For property relationships, we should read as a “Reference Material” can be a “Irradiated Reference Material”, that is, “Irradiated” **is a property** or a **state** of the “Reference Material”. Even though, at first sight, some classes sounds strange, the overall ontology consistency is checked using appropriate tools, some of them included in the Protégé software.

In the Figure 3, it’s worth noting that the top class name was chosen as “Analysis Object” meaning the sample that is being analyzed. This was necessary because “Sample”, “Aliquot” and “Irradiated Aliquot” are all classes at the same hierarchy level; that means, for instance, that “Aliquot” **is not** a type of “Sample”. Rather, “Aliquot” is part of a “Sample”. In this same figure, we have “Treated Sample” as being a property or quality of “Aliquot” or of “Sample”.

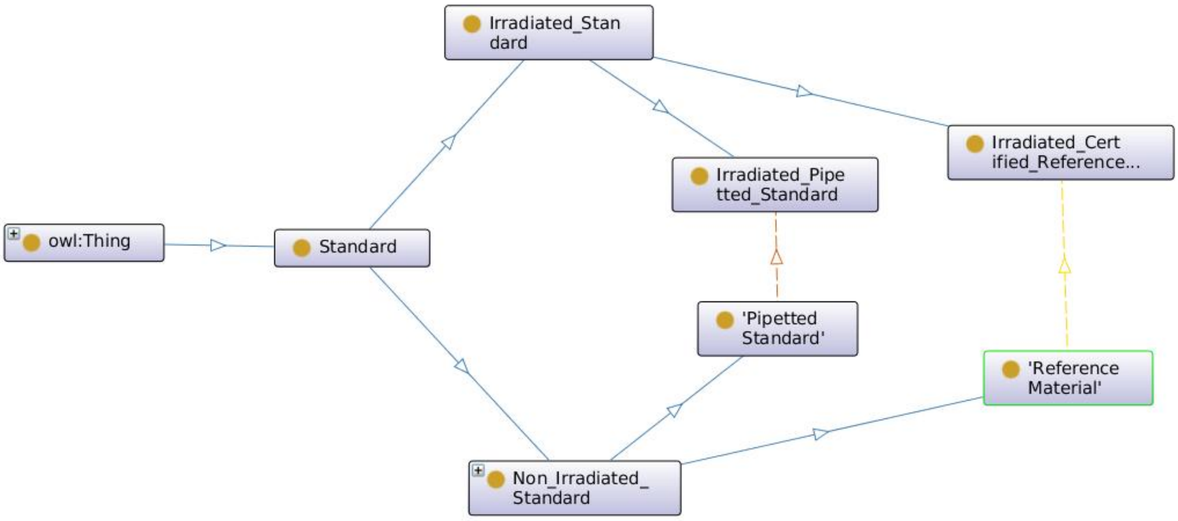


Figure 2: Standards: Certified Reference Material and Pipetted Standard classes

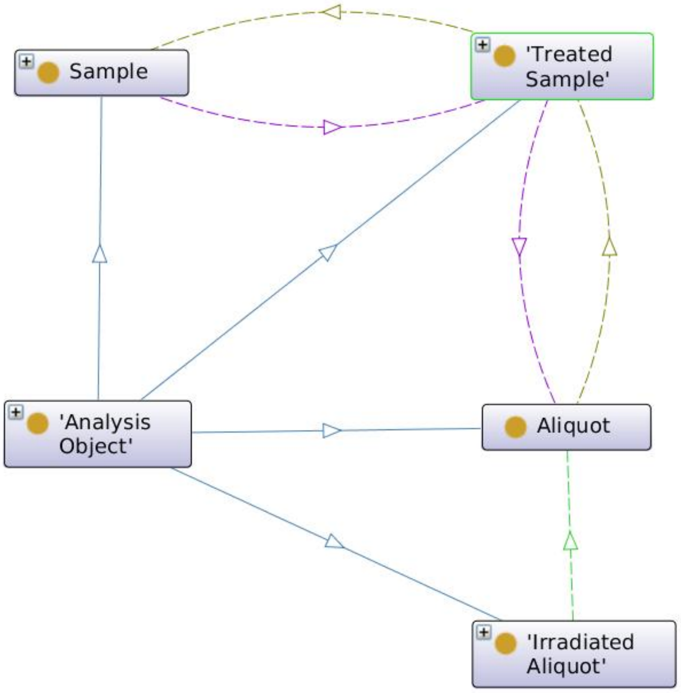


Figure 3: The Object being analyzed, i.e., sample, aliquot, irradiated sample and aliquot classes

2.3. Repository prototype development

The repository prototype is currently being developed using open-source software. The programming language of choice is Python 3 together with the Django 2.1 web development framework. Python is a multiplatform language, object oriented, easy to learn but still powerful in its features [11]. Django is a high-level Python Web framework that encourages rapid development and clean, pragmatic design [12].

PostgreSQL is the relational database choosen for the project. PostgreSQL is a powerful, open source object-relational database system with over 30 years of active development that has earned it a strong reputation for reliability, feature robustness, and performance [13].

The repository prototype basic funcionality includes storing and retrieving NAA research data of the NAA laboratory. User friendly interfaces will allow the users to enter all data and metadada about their research including auxilary files, like spreadsheets, spectra file, or documents with the concentration results.

The architecture for the integration of the ontology with the relational database that stores the data from the NAA researches is currently in development. This architecture aims to integrate the database (relational type) with the user interface, using an intermediate layer composed of the RDFlib framework that will work in conjunction with the generated ontology [14].

Part of this integration is begin developed using the SPARQL (a recursive acronym), which denotes the SPARQL Protocol and the RDF Query Language, recommended by the World Wide Web Consortium (W3C) [15]. The core of SPARQL is composed of simple queries in the form of basic graph patterns made to retrieve and manipulate data in RDF format. A basic graph pattern matches a subgraph of the RDF data when RDF terms from that subgraph may be substituted for the variables and the result is RDF graph equivalent to the subgraph. Data stored in triplets can be manipulated by the SPARQL language through SPARQL Endpoints.

Queries with SPARQL resembles ones done with SQL; two example are shown following:

- Search for samples that have gone through the drying process.

```
PREFIX owl:   <http://www.w3.org/2002/07/owl#>
PREFIX rdf:   <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs:  <http://www.w3.org/2000/01/rdf-schema#>
PREFIX inaa1: <http://www.ipen.br/ontologies/inaa#>
SELECT ?treated_aliquot
  WHERE {?treated_aliquot inaa1:HasGoneTreatmentMethod
      inaa1:Drying
}
```

And the result is:

```
inaa1:ATBPERN01
inaa1:ATBPERN02
```

- Search for individuals belonging to classes and their respective types.

```
PREFIX owl:   <http://www.w3.org/2002/07/owl#>
PREFIX rdf:   <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs:  <http://www.w3.org/2000/01/rdf-schema#>
PREFIX inaa1: <http://www.ipen.br/ontologies/inaa#>
SELECT *
  WHERE {?individual rdf:type ?type .
      OPTIONAL { ?type rdfs:subClassOf ?class }
}
ORDER BY ?class
```

  And the result is:

```
inaa1:ATAAGUA01   Environmental_Treated_Aliquot   Aliquot
inaa1:ATAAGUA02   Environmental_Treated_Aliquot   Aliquot
```

The integration of SPARQL with Django application software being developed will be done with the RDFlib. For instance, the first query above would be implemented with RDFlib as:

```
import rdflib
g = rdflib.Graph()
# ... add some triples to g somehow ...
g.parse("inaa.rdf")
qres = g.query(
    """SELECT ?treated_aliquot
         WHERE {?treated_aliquot inaa1:HasGoneTreatmentMethod
                inaa1:Drying
         }""")
```

Several typical queries using SPARQL are being implemented and they will be part of the software user interface (UI); with this, users will be able to access the stored data using not only traditional search tools but also a semantic search layer.

3. CONCLUSIONS

The purpose of this study is to develop a semantic layer for a data repository prototype; this semantic layer main component is the ontology for Neutron Activation Analysis domain, whose developement was carried out following a hybrid methodology. The results so far achieved are in good aggreement with those found in the literature concerning semantic technologies. The main task being developed currently is the integration of the ontology into the relational architecture. With this integration, the semantic layer will be fully functional, allowing users to explore its full potential of search and navigation.

ACKNOWLEDGMENTS

Authors are thankful to all personnel of the Neutron Activation Laboratory of CERPq for their time and dedication to help the understanding of the NAA processes.

REFERENCES

1. Jim Gray. *The Fourth Paradigm – Data-Intensive Scientific Discovery*, chapter Jim Gray on e-Science – A Transformed Scientific Method, pages xvii–xxxi. Microsoft Research, Redmond, Washington, 2009.
2. Fabrício Marques. Ciência transparente. *Revista Fapesp*, (218):54–58, 2014.
3. Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, et al. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3:160018, mar 2016.
4. Mauricio B. Almeida and Marcello P Bax. Uma visão geral sobre ontologias: pesquisa sobre definições, tipos, aplicações, métodos de avaliação e de construção. *Ciência da Informação*, 32(3):7–20, 2003.
5. Marília Gabriela Miranda Catharino. Análise de mercúrio e selênio em materiais biológicos pelo método de análise por ativação com nêutrons. Master's thesis, Instituto de Pesquisas Energéticas e Nucleares - IPEN/CNEN-SP, São Paulo, 2002.
6. Faiez Gargouri. *Ontology Theory, Management and Design: Advanced Tools and Models (Premier Reference Source).* Information Science Reference; First edition, Hershey, first edit edition, 2010.
7. Natalya F Noy and Deborah L. McGuinness. Ontology development 101: A guide to creating your first ontology. *Stanford Knowledge Systems Laboratory*, page 25, 2001.
8. Sandro Rautenberg, José L. Todesco, and Fernando A. O. Gauthier. Processo de desenvolvimento de ontologias: uma proposta e uma ferramenta. *Revista Tecnologia*, 30(1):133–144, jun 2016.
9. Karin Breitman. *Web Semântica - A Internet do Futuro.* LTC - Livros Técnicos e Científicos Editora S.A., Rio de Janeiro, RJ, 1^{a} edition, 2005.
10. Mark A. Musen. The Protégé Project: A look back and a look forward. *AI Matters*, 1(4):4–12, jun 2015.
11. Python Programming Language. `http://www.python.org`.
12. Django python web-framework. `http://www.djangoproject.com`.
13. Postgresql. `http://www.postgresql.org`.
14. Rdflib Python package to working with rdf. `https://rdflib.readthedocs.io`.
15. W3C. SPARQL Protocol And RDF Query Language. `https://www.w3.org/TR/rdf-sparql-query`, 2008.