Online Police Station, a cutting edge service against cybercrime

C. Aliprandi[1], L. Lotti[2], F. Neri[1] & G. Sanna[2]
[1]*Synthema, Italy*
[2]*Minister of the Interior, Italian State Police, Postal and Communications Police, Italy*

Abstract

This paper describes a content enabling system that provides deep semantic search and information access to large quantities of distributed multimedia data for both experts and the general public. It provides a language independent search and dynamic classification features for a broad range of data collected from several sources in a number of culturally diverse languages. This system is part of the Online Police Station, launched by the Italian Minister of the Interior in 2006. The Online Police Station uses a virtual reality interface to provide general information and online assistance. Citizens can download forms, make complaints, receive advice and/or report events of an illegal nature. Police specialists can monitor criminal trends to ensure that responses are appropriately focused, and that scarce resources are more effectively employed against criminality. Online Police Station was voted as the *Most inspiring good practice for creative solutions to common challenges*, during the last *European eGovernment Awards 2007*.

Keywords: Online Police Station, cyber crime, pornography, e-mail abuse, online gambling and betting, open source intelligence, focused crawling, natural language processing, morphological analysis, syntactic analysis, functional analysis, supervised clustering, unsupervised clustering.

1 Online Police Station

The availability of a huge amount of data in the open sources information channels leads to the well-identified modern paradox: an overload of information means, most of the time, a no usable knowledge. The process of accessing all

WIT Transactions on Information and Communication Technologies, Vol 40, © 2008 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA080231

these raw data, heterogeneous both for type (web pages, claims, crime reports), source (Internet/Intranet, database, etc), protocol (HTTP/HTTPS, FTP, GOPHER, IRC, NNTP, etc) and language used, transforming them into information, is therefore inextricably linked to the concepts of textual analysis and synthesis, hinging greatly on the ability to master the problems of multilinguality.

Therefore, the Italian Minister of the Interior has adopted a multilingual indexing, searching and clustering system, designed to manage huge sets of data collected from different and geographically distributed information sources, in order to overcome information overload. OPS, the first world's online police station, was launched on February 2006, the 15th. OPS uses a virtual reality interface to provide general information and online assistance to citizens and policemen and is made available to the public at the URL www.commissariatodips.it. The commissariatodips.it website allows citizens to seek general information, download forms and make complaints about computer crimes. In addition, it is possible to receive valuable advice, interact with experts and report illegal conduct and events on the web. As the police station is the point of reference in the event of difficulties, OPS was created to be a point of reference on the web. Furthermore the system assists specialists on monitoring criminal trends to ensure that responses are appropriately focused, and that scarce resources are more effectively used against criminality.

The ICT security area is the actual operations centre offering web users professional and specialized support, providing them with useful information and advice on how to surf the web safely. This is a specific area where users can delve into the subject, submit complaints and make reports online, and discuss with police experts in specific forums. With this innovative portal, the first and so far the only one in the world, citizens can give their contribution to step up security levels. They can count on and inform the Postal and Communications Police experts about dangerous aspects and situations regarding the Internet. The increasingly closer contacts with citizens have also disclosed new forms of crime. This service has lightened the workload of conventional police stations, and more officers can now be assigned to other tasks. So far results are excellent, especially considering the increasing number of both visitors to the OPS website and the requests of the services it offers.

1.1 State of art

Current-generation information retrieval (IR) systems excel with respect to scale and robustness. However, if it comes to deep analysis and precision, they lack power. Users are limited by keywords search, which is not sufficient if answers to complex problems are sought. This becomes more acute when knowledge and information are needed from diverse linguistic and cultural backgrounds, so that both problems and answers are necessarily more complex. Developments in the IR have mostly been restricted to improvements in link and click analysis or smart query expansion or profiling, rather than focused on a deeper analysis of text and the building of smarter indexes.

Traditionally, text and data mining systems can be seen as specialized systems that convert more complex information into a structured database, allowing people to find knowledge rather than information. For some domains, text mining applications are well-advanced, for example in the domains of medicine, military and intelligence, and aeronautics [1].

In addition to domain-specific miners, general technology has been developed to detect Named Entities [2], co-reference relations, geographical data [3], and time points [4].

The field of knowledge acquisition is growing rapidly with many enabling technologies being developed that eventually will approach Natural Language Understanding (NLU). Despite much progress in Natural Language Processing (NLP), the field is still a long way from language understanding. The reason is that full semantic interpretation requires the identification of every individual conceptual component and the semantic roles it play. In addition, understanding requires processing and knowledge that goes beyond parsing and lexical lookup and that is not explicitly conveyed by linguistic elements. First, contextual understanding is needed to deal with the omissions. Ambiguities are a common aspect of human communication. Speakers are cooperative in filling gaps and correcting errors, but automatic systems are not. Second, lexical knowledge does not provide background or world knowledge, which is often required for non-trivial inferences.

Any automatic system trying to understand a simple sentence will require – among others – accurate capabilities for Named Entity Recognition and Classification (NERC), full Syntactic Parsing, Word Sense Disambiguation (WSD) and Semantic Role Labeling (SRL) [5].

Current baseline information systems are either large-scale, robust but shallow (standard IR systems), or they are small-scale, deep but ad hoc (Semantic-Web ontology-based systems). Furthermore, these systems are maintained by experts in IR, ontologies or language-technology and not by the people in the field. Finally, hardly any of the systems is multilingual, yet alone cross-lingual and definitely not cross-cultural.

The next table gives a comparison across different state-of-the-art information systems, where we compare ad-hoc Semantic web solutions, wordnet-based information systems and tradition information retrieval with OPS [6].

Table 1: Comparison of semantic information systems.

Features	Semantic web	Wordnet-based	Traditional Information retrieval	OPS
Large scale and multiple domains	NO	YES	YES	YES
Deep semantics	YES	NO	NO	YES
Automatic acquisition/indexing	NO	YES/NO	YES	YES
Multi-lingual	NO	YES	YES	YES
Cross-lingual	NO	YES	NO	YES
Data and fact mining	YES	NO	NO	YES

This system bridges the gap between expert technology and end-users that need to be able to use the complex technology.

2 The logical components

The system is built on the following components:
- a Crawler, an adaptive and selective component that gathers documents from Internet/Intranet sources,
- a Lexical system, which identifies relevant knowledge by detecting semantic relations and facts in the texts,
- a Search engine that enables Functional, Natural Language and Boolean queries,
- a Classification system which classifies search results into clusters and sub-clusters recursively, highlighting meaningful relationships among them.

2.1 The crawler

In any large company or public administration the goal of aggregating contents from different and heterogeneous sources is really hard to be accomplished. Searchbox is a multimedia content gathering and indexing system, whose main goal is managing huge collections of data coming from different and geographically distributed information sources. Searchbox provides a very flexible and high performance dynamic indexing for content retrieval [7, 8].

In Searchbox, the *gatherer* is the coordinator of a pool of agents whose task is to acquire new data from an information source, as soon as it is available. For instance, a noticeable example of a gathering agent is the focused Web crawler, which starts form a set of initial Web pages – the seeds – and performs intelligent navigation on the basis of appropriate classifiers. The gathering activities of Searchbox, however, are not limited to the standard Web, but operate also with other sources like remote databases by ODBC, Web sources by FTP-Gopher, Usenet news by NNTP, WebDav and SMB shares, mailboxes by POP3-POP3/S-IMAP-IMAP/S, file systems and other proprietary sources.

The *renderer* is a central component in the Searchbox architecture. Searchbox indexing and retrieval system does not work on the original version of data, but on the "rendered version". Any piece of information (e.g. a document) is then processed to produce a set of features using appropriate algorithms. For instance, the features extracted from a portion of text might be a list of keywords/lemmas/concepts, while the extraction of features from a bitmap image might be extremely sophisticated. Even more complex sources, like video, might be suitably processed so as to extract a textual-based labeling, which can be based on both the recognition of speech and sounds. All extracted features are then compiled in an internal XML format and passed to the indexing module. The extraction process of the renderer component is done by a pipeline of plug-ins, which provides the compilation of the final XML representation.

The *indexer* creates the index of the collection of information gathered from multiple sources, while the querying module offers a complete query language

for retrieving original contents, wading through millions of documents. The index is fully dynamic in the sense that any indexed content is almost-immediately available for queries. This is a crucial feature when the system is used on highly dynamic sources.

Searchbox indexer module can manage any feature that a specific renderer plug-in is able to extract from the original raw content. All of the extracted and indexed features can be combined in the query language which is available in the user interface. Searchbox provides default plug-ins to extract text from most common types of documents, like HTML, XML, TXT, PDF, PS and DOC. Other formats can be supported using specific plugins. Finally, a multilevel cache is available: the possibility to "historicize" different versions of the same document is a relevant practical feature, which turns out to be especially interesting for the implementation of the watch and alert concepts, when managing tons of documents.

2.1.1 Focused crawling

Focused crawling [9] aims to crawl only the subset of the Web pages related to a specific category. The major problem in focused crawling is performing the appropriate credit assignment to different documents along a crawl path, such that short-term gains are not pursued at the expense of less-obvious crawl paths that ultimately yield larger sets of valuable pages. To address this problem the focused crawling algorithm builds a model for the context within which topically relevant pages occur on the Web. This algorithm shows significant performance improvements in crawling efficiency over standard focused crawling. In fact, the credit assignment can be significantly improved by equipping the crawler with the capability of modelling the context within which the topical materials is usually found on the Web. Such a context model has to capture typical link hierarchies within which valuable pages occur, as well as describe off-topic content that co-occurs in documents that are frequently closely associated with relevant pages. The general framework and the specific implementation of such a context model are called Context Graph. It has a rapid and efficient initialization phase, being suitable for real-time services. The Context Focused Crawler (CFC) uses the limited capability of search engines – like *Google* – to allow users to query for pages linking to a specified document. This data can be used to construct a representation of pages that occur within a certain link distance (defined as the minimum number of link traversals necessary to move from one page to another) of the target documents. This representation is used to train a set of classifiers, which are optimized to detect and assign documents to different categories based on the expected link distance from the document to the target document. During the crawling stage the classifiers are used to predict how many steps far from a target document the current retrieved document is likely to be. This information is then used to optimize the search. There are two distinct stages to using the algorithm when performing a focused crawl session:

(1) An initialization phase when a set of context graphs and associated classifiers are constructed for each of the seed documents.
(2) A crawling phase that uses the classifiers to guide the search, and performs online updating of the context graphs.

2.2 The lexical system

This component is intended to identify relevant knowledge from the whole raw text, by detecting semantic relations and facts in texts. Concept extraction and text mining are applied through a pipeline of linguistic and semantic processors that share a common ground and a knowledge base. The shared knowledge base guarantees a uniform interpretation layer for the diverse information from different sources and languages. The extracted knowledge and information will be indexed by Searchbox, that can handle fast semantic search. The automatic linguistic analysis of the textual documents is based on Morphological, Syntactic, Functional and Statistical criteria. Recognizing and labeling semantic arguments is a key task for answering *Who*, *When*, *What*, *Where*, *Why* questions in all NLP tasks in which some kind of semantic interpretation is needed, like Information Extraction, Question Answering, Summarization.

At the heart of the lexical system is the McCord's theory of Slot Grammar [10]. A slot, explains McCord, is a placeholder for the different parts of a sentence associated with a word. A word may have several slots associated with it, forming a *slot frame* for the word. In order to identify the most relevant terms in a sentence, the system analyzes it and, for each word, the Slot Grammar parser draws on the word's slot frames to cycle through the possible sentence constructions. Using a series of word relationship tests to establish the context, the system tries to assign the context-appropriate meaning (sense) to each word, determining the meaning of the sentence. Each slot structure can be partially or fully instantiated and it can be filled with representations from one or more statements to incrementally build the meaning of a statement. This includes most of the treatment of coordination, which uses a method of 'factoring out' unfilled slots from elliptical coordinated phrases. The parser – a bottom-up chart parser – employs a parse evaluation scheme used for pruning away unlikely analyses

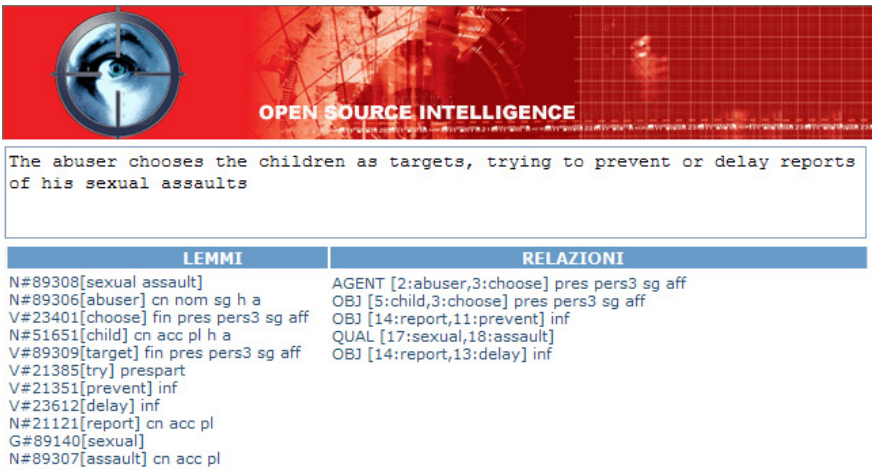

Figure 1: Lexical analysis.

during parsing as well as for ranking final analyses. By including semantic information directly in the dependency grammar structures, the system relies on the lexical semantic information combined with functional relations. Beside Named Entities, locations, time-points, etc, it detects relevant information like noun phrases which comply with a set of pre-defined morphological patterns and whose information exceeds a threshold of significance [13].

The detected terms are then extracted, reduced to their *Part Of Speech*(NOUN, VERB, ADJECTIVE, ADVERB, etc.) *and Functional*(AGENT, OBJECT, WHERE, CAUSE, etc.) tagged base form [12]. Once referred to their synset – namely a group of (near) synonyms - inside the domain dictionaries and knowledge bases, they are used as documents metadata [12–14].

Each synset denotes a concept that can be referred to by its members. Synsets are interlinked by means of semantic relations, such as the super-subordinate relation (hypernymy/hyponymy), the part-whole relation (holonomy/meronymy), antonymy, and several lexical entailment relations. The resultant semantic network allows the human users and automatic systems to navigate the lexicon, identify meaning-related words and concepts, and quantify the degree of their similarity. The domain resources editor lets specialists in the field modify and extend the domain level of the ontology. So the system can be easily maintained and kept up to date.

2.3 Functional navigation

Users can search and navigate by roles, exploring sentences and documents by the functional role played by each concept/lemma, as shown in Figure 2. Users can navigate on the relations chart by simply clicking on nodes or arches, expanding them and having access to set of sentences/documents characterized by the selected criterion.

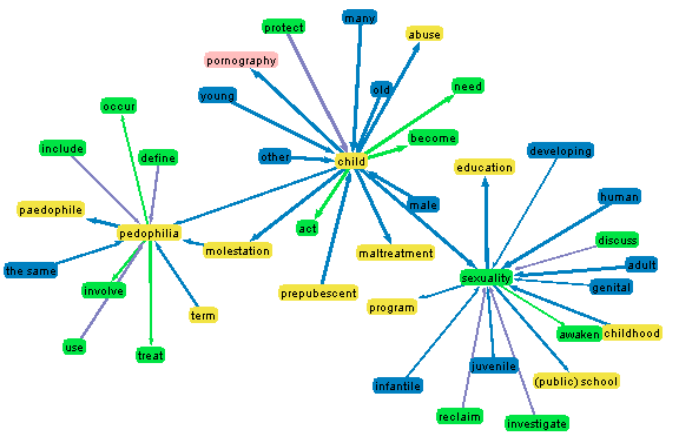

Figure 2: Functional search and navigation.

This can be considered a visual investigative analysis component specifically designed to bring clarity to complex investigations. It automatically enables investigative information to be represented as visual elements that can be easily analyzed and interpreted. Functional relationships – *Agent, Action, Object, Qualifier, When, Where, How* – among human beings and organizations can be searched for and highlighted, pattern and hidden connections can be instantly revealed to help investigations, promoting efficiency into investigative teams. Should human beings be cited, their photos can be shown by simple clicking on the related icon.

2.4 The search

Users can search documents by query in Natural Language, expressed using normal conversational syntax, or by keywords combined by Boolean operators. Reasoning over facts and ontological structures makes it possible to handle diverse and more complex types of questions. Traditional Boolean queries in fact, while precise, require strict interpretation that can often exclude information that is relevant to user interests. So this is the reason why the system analyzes the query, identifying the most relevant terms contained and their semantic and functional interpretation. By mapping a query to concepts and relations very precise matches can be generated, without the loss of scalability and robustness found in regular search engines that rely on string matching and context windows. The search engine returns as result all the documents which contain the query concepts/lemmas in the same functional role as in the query, trying to retrieve all the texts which constitute a real answer to the query.

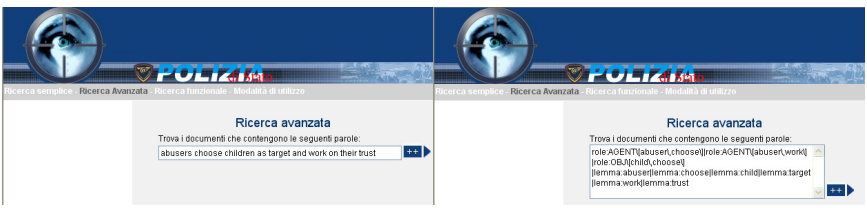

Figure 3: Natural language query and its functional and conceptual expansion.

Results are then displayed and ranked by relevance, reliability and credibility.

Conceptual and lexical descriptors can be exported to I2 Analyst's Notebook®, or to Microsoft® Excel.

2.5 The clustering system

The automatic classification of results is made by *TEMIS Insight Discoverer Categorizer* and *Clusterer*, fulfilling both the Supervised and Unsupervised Classification schemas. The application assigns texts to predefined categories and dynamically discovers the groups of documents which share some common traits.

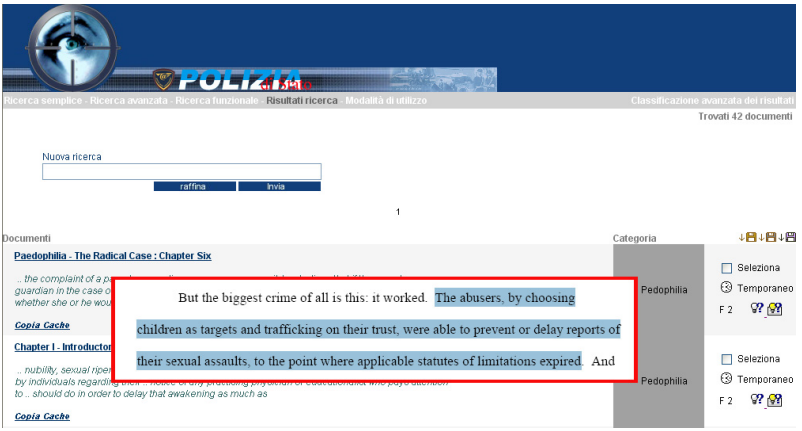

Figure 4: Search results.

2.5.1 Supervised clustering

The categorization model was created during the learning phase, on representative sets of training documents focused on cyber crime (pedo-pornography, telephony, phising and hacking). The bayesian method was used as the learning method: the probabilist classification model was built on around 1.000 documents. The overall performance measures used were *Recall* (number of categories correctly assigned divided by the total number of categories that should be assigned) and *Precision* (number of categories correctly assigned divided by total number of categories assigned): in our tests, they were 75% and 80% respectively.

2.5.2 Unsupervised clustering

Result documents are represented by a sparse matrix, where lines and columns are normalized in order to give more weight to rare terms. Each document is turned to a vector comparable to others. Similarity is measured by a simple cosines calculation between document vectors, whilst clustering is based on the K-Means algorithm, chosen for its simplicity and speed when applied to large datasets. The application provides a visual summary of the clustering analysis. A map shows the different groups of documents as differently sized bubbles (the size depends on the number of documents contained) and the meaningful correlation among them as lines drawn with different thickness. Users can search inside topics, project clusters on lemmas and their functional links.

3 Conclusions

The main results obtained since the launch of the OPS can be summarized as follow:

- Less work for the conventional Police Stations, more effectively employed against criminality.

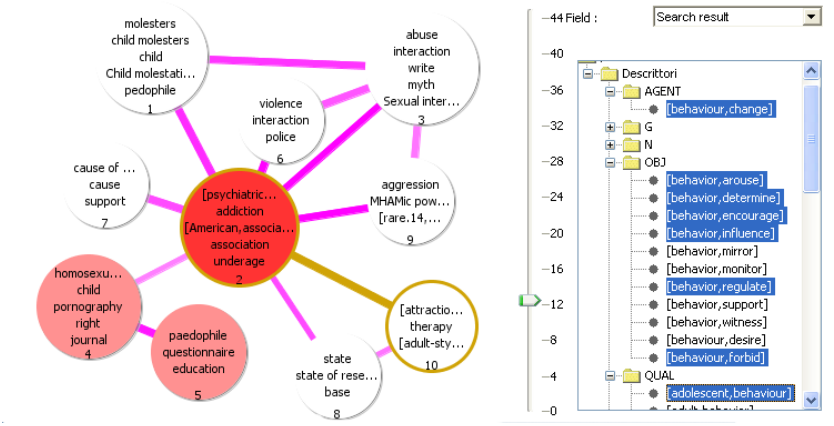

Figure 5: Thematic map, functional search and projection inside topics.

- A new concept: the citizens in the middle of the Institutions. In fact, the Online Police Station reaches them in their office, home and so on. Therefore the citizens' trust in the institutions has increased and now they can browse the Net more safely.
- More trust in the institutions, more crimes reported, ready knowledge of new crimes on the Net, more security of the citizens surfing the Net, full knowledge of social phenomena and new kind of crimes (bullysm, etc).

OPS was voted as the *Most inspiring good practice for creative solutions to common challenges*, during the last *European eGovernment Awards*.

References

[1] Grishman, R., Sundheim, B., Message Understanding Conference – 6: A Brief History. In: Proceedings of the 16th International Conference on Computational Linguistics (COLING), I, Kopenhagen, 1996, 466–471.

[2] Hearst, M., *Untangling Text Data Mining*, *ACL'99*. University of Maryland, June 20–26, 1999.

[3] Miller, H.J., Han, J., *Geographic Data Mining and Knowledge Discovery*, CRC Press, 2001.

[4] Li Wei, Eamonn Keogh, *Semi-Supervised Time Series Classification*, SIGKDD, 2006.

[5] Carreras, X., Màrquez, L., *Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling*. In CoNLL-2005, Ann Arbor, MI USA, 2005.

[6] Vossen, P., Neri, F. et al., *KYOTO: A System for Mining, Structuring, and Distributing Knowledge Across Languages and Cultures, Proceedings of GWC 2008, The Fourth Global Wordnet Conference,* Szeged, Hungary, January 2008, 22–25.

[7] Baldini, N., Bini, M., *Focuseek searchbox for digital content gathering*, AXMEDIS 2005 – 1st International Conference on Automated Production

of Cross Media Content for Multi-channel Distribution, Proceedings Workshop and Industrial pp. 24–28.

[8] Baldini, N., Gori, M., Maggini, M., *Mumblesearch: Extraction of high quality Web information for SME*, 2004 IEEE/WIC/ACM International Conference on Web Intelligence.

[9] Diligenti, M., Coetzee, F. M., Lawrence, S., Giles, C. L., Gori, M., *Focused Crawling Using Context Graphs*, Proceedings of 26th International Conference on Very Large Databases, VLDB, pp. 527–534, September 2000, 10–12.

[10] McCord, M. C., *Slot Grammar: A System for Simpler Construction of Practical Natural Language Grammars* Natural Language and Logic 1989: 118–145. McCord, M. C., *Slot Grammars*, American Journal of Computational Linguistics 6(1): 31–43 (1980).

[11] Marinai, E., Raffaelli, R., *The design and architecture of a lexical data base system*, COLING'90, Workshop on advanced tools for Natural Language Processing, Helsinki, Finland, August 1990, 24.

[12] Cascini, G., Neri, F., *Natural Language Processing for Patents Analysis and Classification*, ETRIA World Conference, TRIZ Future 2004, Florence, Italy.

[13] Neri, F., Raffaelli, R., *Text Mining applied to Multilingual Corpora*, Knowledge Mining: Proceedings of the NEMIS 2004 Final Conference, Springer Verlag Pub., Spiros Sirmakessis Ed., ISBN-13: 978–3540250708.

[14] Baldini, N., Neri, F., *A Multilingual Text Mining based content gathering system for Open Source Intelligence*, IAEA International Atomic Energy Agency, Symposium on International Safeguards: Addressing Verification Challenges, Wien, Austria, IAEA-CN-148/192P, Book of Extended Synopses, pp. 368–369, October 2006, 16–20.