NeuroExtract: Facilitating Neuroscience-oriented Retrieval from Broadly-focused Bioscience Databases Using Text-based Query Mediation

CHIQUITO J. CRASTO, PHD, PETER MASIAR, MS, PERRY L. MILLER, MD, PHD

Abstract
This paper describes NeuroExtract, a pilot system which facilitates the integrated retrieval of Internet-based information relevant to the neurosciences. The approach involved extracting descriptive metadata from the sources using domain-specific queries; retrieving, processing, and organizing the data into structured text files; searching the data files using text-based queries; and, providing the results in a Web page along with descriptions to entries and URL links to the original sources. NeuroExtract has been implemented for three bioscience resources, SWISSPROT, GEO, and PDB, which provide neuroscience-related information as sub-topics. We discuss several issues that arose in the course of NeuroExtract’s implementation. This project is a first step in exploring how this general approach might be used, in conjunction with other query mediation approaches, to facilitate the integration of many Internet-accessible resources relevant to the neurosciences.

Introduction
This paper describes an approach that facilitates the integrated retrieval of Internet-based information relevant to the neurosciences. There are a growing number of genomic and proteomic databases that contain large amounts of data, only a fraction of which are directly relevant to the neurosciences. This paper describes one approach to trying to make neuroscience-specific data retrieval more flexible and easily integrated.

The approach involves the following steps:
• Extracting the descriptive metadata using domain-specific queries;
• Processing and organizing the extracted data into a structured text file containing metadata from the entry, relevant keywords for searching, and information that links to the original source;
• Searching the extracted metadata;
• Creating search results and providing links to the metadata source.

Background
In addition to neuroscience-specific online data sources, there also exist a growing number of national/international bioscience sources that house neuroscience information as a subset of much broader sets of available data. In order to access such structured information stored in a database,1–3 a user has to enter a structured query that is processed by a server-side algorithm before the requested information is retrieved. This paper discusses a query integration system in the neuroscience domain. Cheshire and PESTO,4 which variably query databases, unstructured text and imaging data have also been developed; but these systems unlike NeuroExtract do not explore textual extraction as a vehicle for query mediation.

A variety of approaches have been undertaken to explore the integration and interoperation of neuroscience data. The Neuroscience Database Gateway (NDG) (http://ndg.sfn.org/) was created as a pilot project for the Society for Neuroscience as a repository of neuroscience-related Internet-based information sources. The NDG categorizes these sources in a variety of ways, e.g., experimental data, neuroscience knowledge bases, software tools, informatics resources, with recent additions of sources containing proteomics and genomics information related to neuroscience. The Neuro-
science Information Gateway (NIF) has been recently developed (building, in part, upon the NDG) with the goal of providing a comprehensive listing of online resources related to neuroscience (http://neurogateway.org/catalog/). Work is ongoing to create and extend the list of neuroscience concepts and ontologies that can be mapped to sources described in the NIF. The creation of the NDG and NIF as sources for neuroscience information arose from the need for interoperability and information sharing. This exchange of data between laboratories and databases aims to enhance data availability to researchers and end-users and avoid redundancies due to storage of similar information in more than one database.

One form of interoperation involves query mediation, the integrated querying of multiple databases in a coordinated fashion. BIRN and Query Integrator System (QIS) involve executing queries directly upon a set of federated databases. The NeuroExtract approach differs in its use of textual descriptions extracted from a set of databases to facilitate the query mediation process.

**Design of Pilot Neuroextract System**

**NeuroExtract’s Three Pilot Knowledge Sources**

To create a pilot text-searchable repository of neuroscience information, three genomic/proteomic resources were chosen.

- **SWISSPROT** (http://ca.expasy.org/sprot/): is a comprehensive resource for all identified proteins. It provides descriptors for proteins, links to additional information for the protein and tools to process the information, in terms of sequence and structural analyses.

- **Gene Expression Omnibus (GEO)** (http://www.ncbi.nlm.nih.gov/geo/): stores gene expression data from microarray experiments. GEO accepts MIAME (Minimum Information About a Microarray Experiment) compliant data and contains interfaces to query, search, and retrieve microarray data.

- **Protein Database (PDB)** (http://www.rcsb.org/pdb): is a repository of information related to the results of experimentally and theoretically derived structures of pro-
proteins, DNA, and protein-DNA, complexes. PDB also provides abstracts of the publication related to the protein structures. Each abstract is linked to a set of keywords that allow for the easy indexing and querying of PDB entries.

Extracting Relevant Data
SWISSPROT, GEO, and PDB are not primarily neuroscience databases. They do, however, contain entries related to neuroscience. Separate word searches on keywords “brain” and “central nervous system” on all three databases resulted in: 32,985 unique entries from SWISSPROT, 507 entries from GEO, and 425 entries from PDB. Individual algorithms to process each of the query-result files were developed.

- The results of the SWISSPROT query were downloaded in a single text file. The text for each entry was processed by using two-letter identifier tags; for example, “AC” denotes the SWISSPROT entry number and “RT” is a pointer to the title of the entry.
- GEO query results are available as html files or text files. We processed the text-formatted entries, since the information provided therein was sufficient for the data file creation. Extraction of relevant keywords and concepts was accomplished by comparing the text of the entry to a neuroscience keyword list as described in the Discussion section.
- PDB entries are available as downloadable HTML. A list of relevant PDB accession IDs was first extracted. A script was then used to automatically download and process each entry. This script dynamically generates a URL containing the PDB accession IDs and a link to the PUBMED abstract within PDB.

For the present case study, we have taken a semi-automated approach to extracting data from the three databases. If NeuroExtract were to become an operational system, this extraction process would need to be automated to keep pace with the growth in bioscience databases.

Creating NeuroExtract’s Data Files
The information downloaded from SWISSPROT, GEO, and PDB was processed to create a structured text data file that could be accessed using a single text-based retrieval program. A glossary of neuroscience terms (keywords) was used to process each entry’s title, summary, and abstract. Also included in the text file were links to the original source and to the biomedical literature.

The following is a sample of part of a single line in a data file generated as a result of pre-processing the results of an entry in PDB.

1BTN|PUB7588597|Structure of the binding site for inositol phosphates in a PH domain|Amino Acid Sequence|Binding Sites|Blood Proteins|Cell Membrane|Circular Dichroism|Crystallography|X-Ray|Inositol Phosphates|Magnetic Resonance Spectroscopy|Models|Molecular|Molecular Sequence Data|...

The first field denotes the accession ID (here, 1BTN). Fields are separated by the “/” delimiter. The second field indicates the PUBMED ID and title of Abstract for that entry. The “PUB” prefix before the PUBMED entry number allows the query interface script to recognize that the field contains a PUBMED accession ID. The “PUB” prefix is used to account for SwissProt entries, which are often associated with several PUBMED links. Including the title of the article enables free text searching of the title. The remaining fields contain keywords derived 1) by scanning the abstract using a neuroscience concept/keyword list, and 2) by automatically extracting and storing keywords from PDB.

NeuroExtract’s Query Interface and the Display of Query Results
NeuroExtract’s query interface (Fig. 2) prompts the user to choose from a list of neuroscience terms and concepts. The query interface allows free-text keyword matching to identify information found in the articles’ abstracts (which are available for SwissProt and PDB) with May Match (OR), “Must Match (AND)” and “Exclude (AND NOT)” Boolean operators. The results are presented in a two column format: The title of the entry and a link to the source on the left; and, a link to the PUBMED abstract (for SWISSPROT and PDB results) or species name (for a GEO results) on the right. Figure 2 shows a query where the user seeks information about the keyword “motor.” Partial integrated results of this query are shown in Figure 3. The query resulted in one entry from GEO, three from PDB, and 84 from SWISSPROT. The query results are returned for both “motor neuron” and “motor function.” These queries can be easily customized to exclude results related to motor function. If the word “motor” in the drop down list is ANDed with the term “neuron,” no results are obtained for GEO or PDB and 27 results are returned from SWISSPROT. The results of the latter might mean that the keywords “motor” and “neuron” occur in the entry but do not necessarily refer to “motor neuron.” If, on the other hand, the data files are queried such that they only “MUST MATCH” the phrase “motor neuron,” only 17 results are returned from SWISSPROT—all referring to “motor neuron.” Such easy, iterative query customization across multiple databases to obtain focused results is one advantage of the NeuroExtract approach.

NeuroExtract can also be used to recognize relatedness between two terms, one specifically neuroscience-oriented and the other generic. A search using keywords “cerebellum” and “zinc finger” returned ten entries all from SWISSPROT. Every entry was related to zinc-finger protein. Querying only “zinc finger” resulted in eighteen results in PDB and 635 results from SwissProt, all localized in the central nervous system. The lack of results in PDB when both terms were used in a query indicates that zinc finger proteins associated with the cerebellum have not been structurally characterized. A search query with key phrase “zinc finger” at the SwissProt Web site reveals 1,184 entries. Of these, 635 are related to “brain” or “central nervous system.” This shows that the zinc finger protein is found in tissues other than those associated with the nervous system.

Discussion
This section discusses a number of issues that arose in the implementation of the pilot system.

NeuroExtract as an Integrative Tool
NeuroExtract is a neuroscience-oriented query tool that facilitates text searching of bioscience databases by extracting and processing information using lists of specific concepts, keywords, and phrases. Additional information helps to focus a search, especially where generic neuroscience concepts such as...
"neuron" can return too many results. Different scripts have to be written for each knowledge source, but the information extracted is stored in the data files in a uniform format that can be queried using the same script. NeuroExtract enables the user to compare neuroscience information available in different sources at the same time. The user can readily relate information from different sources since they can be presented on one results page.

Our examples, discussed in the previous section, show that though each of the three sources, GEO, SWISSPROT, and PDB, have their own text searching capabilities, NeuroExtract's integrative querying capability can make information from different sources on a single result while enhancing the breadth and depth of the knowledge acquired about the search. Results, with links to the original sources allow the user to access, for example, a BLAST search or a theoretical

**Figure 2.** NeuroExtract's query interface, as described in the text.

**Figure 3.** NeuroExtract results from all three databases based on the query using the keyword "motor."
structure at SWISSPROT, compare theoretical methods with experimental structures at PDB, and discover if gene expression data from other tissues was also part of a microarray experiment that involved the protein in question.

Extracting Neuroscience-oriented Entries from a Bioscience Database

The simple strategy currently employed by NeuroExtract of using very general searches ("brain" or "central nervous system") to extract information is focused on demonstrating functionality on a pilot basis rather than on completeness. The search options available allow the user to focus a search only within this subset of database entries.

To help assess the effectiveness of this simple initial search, we scanned each of the three databases independently using the 71 keywords in the menu of NeuroExtract’s query interface to determine: 1) how many entries were returned for each of those keywords from each database, and 2) whether the returned entries were also found by our general "brain"/CNS search. Table 1 shows the results of this analysis for the 17 keywords that occur most frequently in the data files. The results for the remaining 54 keywords are summarized in the "Other" category.

Table 1 shows that the information extracted from the "brain/CNS" query do not constitute a full neuroscience corpus. As a result, in implementing a system like NeuroExtract, it will be important to analyze each database in detail. It may be necessary to define quite a complex query to allow a comprehensive set of neuroscience entries to be extracted or "cast the net" more widely than necessary, even if it introduced false positives in NeuroExtract’s files. An ideal solution to this "completeness" problem would be for general bioscience database curators and administrators to index their entries and map them to specific domains such as neuroscience.

Extending NeuroExtract to Other Databases and to Other Domains

Our approach is based on the expectation that a bioscience database will have entries that each contains an experimental result (or a set of related experimental results), together with "descriptive metadata" indicating the nature of the experiment that was performed to obtain those results. This descriptive metadata is typically a combination of standardized keywords and free text. To the extent that a bioscience database matches these expectations, we would anticipate that the NeuroExtract approach could be used.

There are a number of ways in which a bioscience database could be adapted to better accommodate integration into the NeuroExtract approach. Desirable features include, 1) facilitating the automated searching biosciences databases and specific sub-domains, 2) allowing downloading results in a standard format such as XML, and 3) including identifying fields of each entry through standardized tags.

Optimizing NeuroExtract Query Performance

Another issue that arises in extending the use of a tool such as NeuroExtract concerns how best to store the extracted data. NeuroExtract currently uses text files for this purpose. Given the exploratory nature of the present pilot project, using text files is reasonable. Currently, NeuroExtract can process most queries in less than 2 seconds.

After pre-processing, the size of data files based was 44 KB for GEO (489 entries), 190 KB for PDB (420 entries), and 11.4 MB for SwissProt (approximately 34,000 entries). Efficiency issues will become important if NeuroExtract were to incorporate many more databases and much larger datasets. It would be important to explore methods in the future to optimize search performance, for example, by storing the extracted data in a relational database.

Summary

This paper describes an approach that facilitates the integrated querying of multiple biosciences databases for data relevant to the neurosciences, using text-based query mediation. The NeuroExtract system was developed to explore the approach and to help highlight the various issues that arise in implementing the approach. This case-study is a first step in exploring how this general approach might be used, in conjunction with other query mediation approaches, to facilitate the integration of many Internet-accessible resources relevant to the neurosciences.

References


Table 1: The Number of Entries Retrieved from the Three Databases Using Different Search Terms

<table>
<thead>
<tr>
<th>Search Terms</th>
<th>PDB</th>
<th>SWISSPROT</th>
<th>GEO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Broad search</td>
<td>418</td>
<td>33781</td>
<td>489</td>
</tr>
<tr>
<td>brain or &quot;central nervous system&quot;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Focused searches</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;spinal cord&quot;</td>
<td>2/0</td>
<td>12/0</td>
<td>99/54</td>
</tr>
<tr>
<td>cerebellum</td>
<td>24/0</td>
<td>3/0</td>
<td>84/65</td>
</tr>
<tr>
<td>neuron</td>
<td>393/62</td>
<td>109/25</td>
<td>144/62</td>
</tr>
<tr>
<td>axons</td>
<td>13/3</td>
<td>8/0</td>
<td>70/18</td>
</tr>
<tr>
<td>hormones</td>
<td>76/0</td>
<td>11/9</td>
<td>15/14</td>
</tr>
<tr>
<td>glia</td>
<td>71/1</td>
<td>12/0</td>
<td>27/13</td>
</tr>
<tr>
<td>neurotransmission</td>
<td>15/0</td>
<td>123/10</td>
<td>19/12</td>
</tr>
<tr>
<td>synapse</td>
<td>29/0</td>
<td>135/8</td>
<td>13/3</td>
</tr>
<tr>
<td>forebrain</td>
<td>0/0</td>
<td>8/0</td>
<td>35/26</td>
</tr>
<tr>
<td>mesencephalon</td>
<td>2/0</td>
<td>0/0</td>
<td>16/10</td>
</tr>
<tr>
<td>myelencephalon</td>
<td>2/0</td>
<td>2/0</td>
<td>21/14</td>
</tr>
<tr>
<td>dopaminergic</td>
<td>0/0</td>
<td>67/1</td>
<td>0/0</td>
</tr>
<tr>
<td>&quot;olfactory receptor&quot;</td>
<td>2/0</td>
<td>488/0</td>
<td>2/1</td>
</tr>
<tr>
<td>motor</td>
<td>24/0</td>
<td>5/4</td>
<td>44/25</td>
</tr>
<tr>
<td>photoreceptor</td>
<td>0/0</td>
<td>0/0</td>
<td>40/24</td>
</tr>
<tr>
<td>granule cell</td>
<td>0/0</td>
<td>0/0</td>
<td>14/11</td>
</tr>
<tr>
<td>purkinje</td>
<td>0/0</td>
<td>0/0</td>
<td>10/4</td>
</tr>
<tr>
<td>other (54 more keywords)</td>
<td>51/0</td>
<td>45/12</td>
<td>485/215</td>
</tr>
</tbody>
</table>

The numbers after the slash indicate how many of the entries retrieved for each focused search were also retrieved by the broad search (brain or "central nervous system").


