Database Techniques for Biological Materials & Methods

Kenneth Baclawski, Robert Futrelle, Natalya Fridman and Maurice J. Pescitelli*
Northeastern University
College of Computer Science
Boston, Massachusetts 02115
{kenb,futrelle,natasha,mjp}@ccs.neu.edu

Abstract

The Biological sciences produce an enormous research literature every year. Research papers are highly structured documents whose content is not captured using the traditional techniques of information retrieval: keywords and flat text. This is especially true of the Materials & Methods section of experimental papers. A great deal of highly structured information is packed into this section. It involves logical and temporal sequences of operations that combine and operate on materials using various instruments and depending on many parameters. We are designing and implementing databases that will allow this complex knowledge to be represented, stored in object-oriented databases and retrieved. We are developing an application of this technology called the Laboratory Notebook. This application is a software system that will contain personal laboratory information as well as have access to databases of Materials & Methods sections drawn from the literature.

Introduction.

Biology is a very large and diverse field. The primary output of the enterprise is its published research literature, which consists of about 600,000 papers every year. The vast majority of these papers report experimental work and do so in a highly structured manner. Our objective is to develop database and text analysis techniques for making this literature more accessible.

This paper describes techniques we are developing to extend current database and information retrieval technology to represent more of the content of research papers. The data structures are developed in a top-down fashion and build on the structures already in the text such as sections, paragraphs, figures, etc.

The Materials & Methods sections of biological research papers are especially well-suited for database construction:

1. The Materials & Methods section is an important one. Biology is a technique-driven experimental science rather than a theory-driven one.
2. This section is easier to analyze than the rest of the paper. It has a more limited (although still large) vocabulary. The language used is more stylized (although still complex) and is primarily descriptive with no arguments, discussion of hypotheses, etc. The other parts of the paper use more complex syntactic and linguistic forms.

We are developing a database for the Materials & Methods sections of biological research papers. Initially, we will be dealing with papers in the field of bacterial chemotaxis. But the system will be designed to be extendible to other branches of molecular biology. In the future, the contents of papers will all be available electronically, so our database can provide a prototype and testing ground for the large databases that could encompass the content of papers.

This paper will begin with a discussion of related work on text analysis and knowledge acquisition. We then introduce the Materials & Methods schema. A *schema* is the design or description of data to be stored in a database. An *object* or *instance* is a data item or structure conforming to a schema and stored in a database. The Materials & Methods schema is represented in a number of different forms, and we describe how we translate between them. We illustrate how the schema would be used in practice by analyzing some text taken from a biological research paper. One proposed application related to the Materials & Methods database is a laboratory notebook system. We sketch some of the features of this proposed system. We end with some conclusions and directions for future work.

Related Work.

Recent technical advances in molecular biology have allowed the generation of an enormous volume of data, which cannot be dealt with by traditional printed pub-

*This material is based upon work supported by the National Science Foundation under Grant No. IRI-9117030.

lications. A second form of electronic data publishing has been developed to make data available in a world-wide network-accessible database, while methods and conclusions continue to be published in traditional publications (Cinkosky *et al.*, 1991). Databases such as GenBank (Burks *et al.*, 1992) and the EMBL data library (Higgins *et al.*, 1992) serve as repositories for DNA sequence data; PIR (Barker *et al.*, 1991) and SWISS-PROT (Bairoch and Boeckmann, 1991) store protein sequence data; and GDB stores human gene mapping information (Pearson *et al.*, 1992). Data in these databases is now accessible for computer analysis. The content of research publications, which describes the methods and conclusions associated with the data, remains inaccessible to sophisticated computer search. Our efforts are directed toward making the contents of publications accessible by applying database techniques to the actual content of research publication.

The results of (Lehnert and Sundheim, 1991) support the claim that systems incorporating natural language processing techniques are more effective than systems based on statistical techniques alone. An example of applying natural language processing techniques in a limited domain to extract the information from flat text is in (Jacobs and Rau, 1990). This system used a combination of bottom-up and top-down parsing to analyze news articles and to retrieve information about financial companies.

Several systems for extraction of information from text use frames or templates that are very similar to the schema described below. The JASPER system is concerned with the extraction of numerical data (corporate earnings, dividends, etc.) (Andersen *et al.*, 1992). They use pattern matching techniques that look for generalized terms that occur reasonably close to other terms and in a certain relation to numbers. The patterns include variables to be matched to the numbers. The extracted information is then inserted in a knowledge frame that holds a highly structured form of the data. This data is then used to generate news stories. The KBIRD system (Finin *et al.*, 1991) collects information in a similar way using Prolog-like patterns. Multiple candidates to fill portions of the template compete by comparing confidence scores based on how the information was discovered. The templates can then form the basis of a query system. Material & Methods sections have a certain similarity to cooking recipes. The Recipe Acquisition System developed at the University of Connecticut (McCartney *et al.*, 1992) attempts to understand recipes and build structured descriptions for them. This project is similar in spirit to ours but has little emphasis on queries.

Description of the Schema.

Materials & Methods sections describe complex recipes by which input materials are ultimately transformed into output measurements. Basically, there are two different types of information presented in a Materials and Methods section. One is the list of initial materials used in the experiment. If one thinks of this section as describing a very complex recipe, then these are the ingredients. The initial materials are defined without explaining how they were produced. They are typically defined by referring to a paper where the production is described or to a laboratory or company where the material was obtained.

Names of materials can be very complex, both conceptually and typographically. Moreover, there is no universally accepted naming scheme for all possible materials. This is primarily due to the rapid pace at which biology, especially molecular biology, is developing. Nevertheless, there are many standardization efforts going on not only to make it easier for biologists to communicate with one another in research papers, but also to serve the needs of the many databases being developed. Part of the difficulty with the nomenclature of biology is the technical orthography used, e.g., Greek letters, italics and superscripts and subscripts. These details are captured in our system using SGML to encode the text, so all database structures, displays, etc., remain faithful to the original notation.

The Materials & Methods Database will not, at least initially, attempt to understand the names of materials in detail. Rather, it will classify materials into broad categories such as protein, plasmid, and so on. Within a category the name will be represented as marked-up text to facilitate matching when doing searches and to make it possible to display the name with a graphic user interface. As the database evolves the categories will be subclassified in a top-down fashion rather than analyzed bottom-up.

The category of a material will be determined both by features of the name and by its context in the paper. A rich lexicon and thesaurus will be developed using both automated and human input.

The second type of information which is in the Materials & Methods section is the description of the processes. Processes consist of a sequence of steps that transform input substances into output substances. The description of a process step can be either elementary or complex. An elementary step has no elaboration into other steps. It is defined entirely by a small set of parameters such as temperature, duration and so on. A complex step may have some overall parameters, but the significant feature of a complex step is that it is elaborated into a process in its own

right. The elaboration can be as small as a single step. For example, "separated by centrifugation," "washed by resuspension," "collected by centrifugation" or "released by incubation." But an elaboration can also be a series of interconnected steps as in the case of the immunoaffinity chromatography step in Figure 3.

Some parameters are common to all the steps and are mentioned often when describing a process, such as temperature and duration of the overall process. However, most steps have their own unique subset of parameters. It is also interesting to note that normally each step mentioned in the paper is associated with a particular piece of equipment. In fact, the Materials & Methods section is closely related to the lab notebooks usually maintained by biologists as they perform their experiments.

Some distinct process names tend to occur in the same context and are regarded as describing the same process steps by biologists. The terms are similar but not full synonyms. Two such terms may be used with different sets of parameters. The term used apparently depends on what the scientist wants to emphasize about a particular step. An example of this is the pair of words *wash* and *resuspend* where the latter word is often used to emphasize the medium in which cells are resuspended.

The fact that there are similar ways of describing process steps has implications for database lookup and retrieval when answering queries. When a user asks about a *washing* step, the system will also have to look for *wash*, *resuspend* and *resuspension* steps.

A process consists of a sequence of discrete steps that are related to one another in two main ways. In the first relationship, process steps have inputs and outputs that are substances. Since these substances are linked with other process steps as inputs and/or outputs, one could say that process steps are linked to each other via substances. We refer to this relationship as the "production/consumption" relationship. The production/consumption relation is often very standard and not explicitly described: the output of one process is the input of the next process. When this assumption is violated the text usually alerts us to it.

The second way that process steps can be related to one another is hierarchical subdivision: a more complex step is subdivided into a sequence of simpler steps. This relationship is called the "elaboration" relationship. The production/consumption relationship interacts with the elaboration relationship. For example, the inputs to a process step that is being elaborated can be used as input to any (and even several) of the substeps. Similarly, the outputs of an elaborated step must arise from substeps.

It is also important to recognize and represent the *information* output of processes (or procedures). Besides the numerical output of instruments such as scintillation counters and scales, there is important qualitative information such as "The RNA fragments were detected by autoradiography." Most biological experiments are conducted with the goal of producing information rather than materials *per se*. Our database schema will reflect this.

Schema Design Considerations

The schema for the Materials & Methods Database will be described in several ways. The various forms are all derived from a schema written in a semantic data model called the *alpha model* developed as part of a larger project in object-oriented and semantic data modeling at Northeastern University (Baclawski *et al.*, 1989). The derivations from the basic schema are done using data model translation tools developed in this project. The use of the alpha model has a number of advantages:

- It furnishes a rigorous schema that can be translated into other models.
- By having a single schema one ensures that all the various forms of the schema are consistent with one another.
- The alpha model is simpler than but still easily translated into structured text languages like SGML.
- Modifying the schema in all its forms is easily accomplished.

The following are some of the many forms in which one can view the schema:

1. The original schema in the alpha model.
2. A graphic diagram which can be viewed at several levels of detail.
3. A knowledge representation using frames.
4. An SGML Document Type Description (DTD). This can be used both for storage and also as a standard for communicating information.
5. An English language explanation. While this is not a rigorous schema, it is essential for communicating with scientists who cannot be expected to learn semantic data modeling or SGML.
6. A standard relational schema in SQL.

In the rest of this report, we discuss parts of the schema using several of the forms mentioned above. We have chosen some representative samples to illustrate the overall approach.

In the following, we use `typewriter type` for formal terms used in the definition of the schema. These are

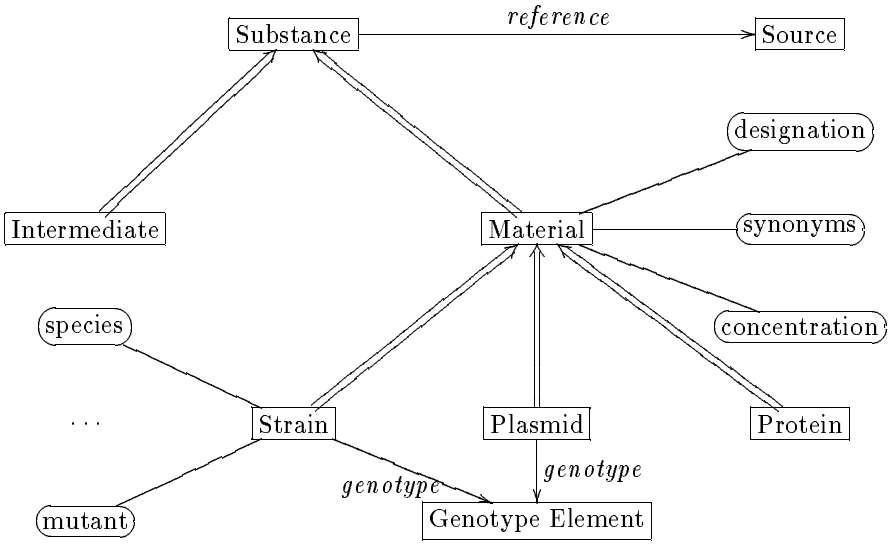

Figure 1: The Materials Schema. This is a part of the Materials schema. Each material is a subclass of *Substance*. There are *Intermediate* substances which generally refer to an unnamed output of some process, which immediately becomes an input to the next step. *Materials* are the substances explicitly named in the paper which have various parameters (*designation, concentration*, etc.). There are various kinds of materials, such as *strains, plasmids* and *proteins*, each of which has its own set of parameters in addition to those of all materials.

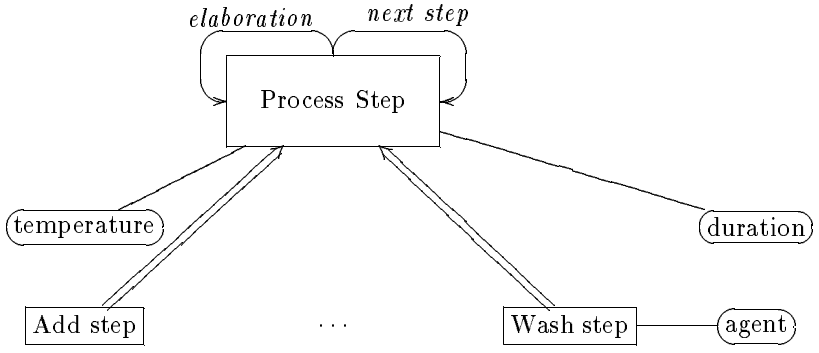

Figure 2: The Process Schema. This figure shows a part of the Process schema. All the process steps are subclasses of the class *Process Step*. A *Process Step* can be an elaboration of another step or the next step in a sequence of steps. Each step may have attributes such as *temperature, duration*, etc. In addition, each step may have its own attributes.

not intended to be biological terms although the terms were chosen to suggest biological concepts.

In Figure 1, the most general category for materials is the `substance`. This type is subclassified into `intermediate` and `material`, which are disjoint. `Intermediate` substances are unnamed substances that serve as the inputs and outputs of processes but are not initial substances. A `material` is uniquely determined by its `designation`. It may have one or more `synonyms`, may have a `concentration` which is a number having some units, etc.

The types `protein`, `plasmid` and `strain` are derived from `material`, that is each material is in one of these categories. Each of the categories has its own set of parameters: for example, `plasmid` may have `feature, replicon` and `parent`. In addition, `plasmid` and `strain` may have some `genotype elements` as its genotype. See Figure 1 for the graphic representation of the schema described above. The description above was automatically generated and then edited to improve the style without affecting the content.

The second important part of the schema is the methods part of the Materials & Methods section, Fig. 2. Each `process step` may have `temperature, duration, input` and `output substances`. It can also be `elaborated` by a sequence of other `process steps`. Every `process step`, whether `elaborated` or not, can have its own (possibly empty) set of attributes. `Process Steps` are subclassified into a large hierarchy of types of step, a few of which are sketched in the diagram.

As mentioned earlier, all of the various forms of the schema are derived from the schema in the alpha model. The following is a small portion of this schema:

```
category substance {
  total disjoint subclassification {
    material, intermediate
  }
  to reference with source;
}
category material {
  string designation;
  unique { designation }
  optional multivalued string synonyms;
  optional float concentration;
  disjoint subclassification {
    strain, plasmid, protein
  }
}
category strain {
  optional multivalued string species;
  optional string parent;
  optional string replicon;
  optional multivalued string feature;
  optional string mutant;
  optional string wild_type_designator;
  association genotype 0..*
    with genotype_element 0..*;
}
category plasmid {
  optional string parent;
  optional string replicon;
  optional multivalued string feature;
  association genotype 0..*
    with genotype_element 0..*;
}
category genotype_element {
  string gene;
  optional string location;
  disjoint subclassification {
    insertion, deletion, fusion
  }
}
```

Part of the SQL representation of the schema is the following:

```
CREATE TABLE material (
  designation CHARACTER VARYING NOT NULL UNIQUE,
  surrogate INTEGER NOT NULL PRIMARY KEY,
  concentration REAL )
CREATE TABLE genotype_element (
  surrogate INTEGER NOT NULL PRIMARY KEY,
  location CHARACTER VARYING,
  gene CHARACTER VARYING NOT NULL )
CREATE TABLE plasmid (
  surrogate INTEGER NOT NULL PRIMARY KEY
    REFERENCES material (surrogate),
  replicon CHARACTER VARYING,
  parent CHARACTER VARYING )
CREATE TABLE material_synonyms (
  material INTEGER NOT NULL
    REFERENCES material (surrogate),
  synonyms CHARACTER VARYING NOT NULL,
  PRIMARY KEY (synonyms, material) )
CREATE TABLE plasmid_genotype (
  plasmid INTEGER NOT NULL
    REFERENCES plasmid (surrogate),
  genotype_element INTEGER NOT NULL
    REFERENCES genotype_element (surrogate),
  PRIMARY KEY (plasmid, genotype_element) )
CREATE TABLE plasmid_feature (
  plasmid INTEGER NOT NULL
    REFERENCES plasmid (surrogate),
  feature CHARACTER VARYING NOT NULL,
  PRIMARY KEY (feature, plasmid) )
```

This would be used to implement the Materials & Methods database using standard relational database technology if object-oriented technology has not advanced sufficiently by the time the database is built.

Analysis of a Paper from the Biological Research Literature.

Consider the following excerpt from a Materials & Methods section from a biology paper (Stock and Stock, 1987). We will analyze this paragraph as data conforming to the schema sketched in the previous section.

> **Immunoaffinity chromatography**. IgG was purified from mouse ascites fluid by DEAE-Affi-Gel Blue (Bio-Rad) chromatography (5) followed by precipitation in 50% ammonium sulfate at 0°C. Purified IgG (5 mg/ml) was dialized against 0.1M sodium bicarbonate, pH 8.5, mixed with Affi-Gel 10 (Bio-Rad) at a ratio of 10 mg of IgG per ml of Affi-Gel 10, and incubated for 12 h at 40°C....

This paragraph is an elaboration on the first part of a particular `process step`: immunoaffinity chromatography. This part deals with the preparation of the column. One of the substeps is itself elaborated into subsubsteps. In Figure 3 we represent this paragraph using knowledge representation frames.

The immunoaffinity chromatography step does not have any parameters of its own, but consists of a series of steps. The first process step in the sequence is *purify*, which has a technique identified with the preposition *by*. The *purify* step is itself elaborated into a series of substeps: *chromatography* and *precipitate*. The description of the *chromatography* technique is not given in this case, instead, we see a reference to a paper where it was described. After the *purify* step, a sequence of steps was carried out, but the intermediate materials produced by these steps are not explicitly mentioned: *dialize, mix* and *incubate*. One can also see the use of parameters for the materials and the steps. Temperature and duration are ones which appear for almost any process. The processes mentioned here also have parameters peculiar to these processes. For example, the step name *precipitation* is followed by the name of the *medium* where the process was carried out. This parameter is associated with this particular process. The same can be said about the *antagonist* for the *dialize* process.

An encoding of the knowledge frames in Figure 3 using an SGML interchange format is shown in Figure 4.

The analysis described above was done by hand, but we are building tools that can automatically generate these knowledge frames.

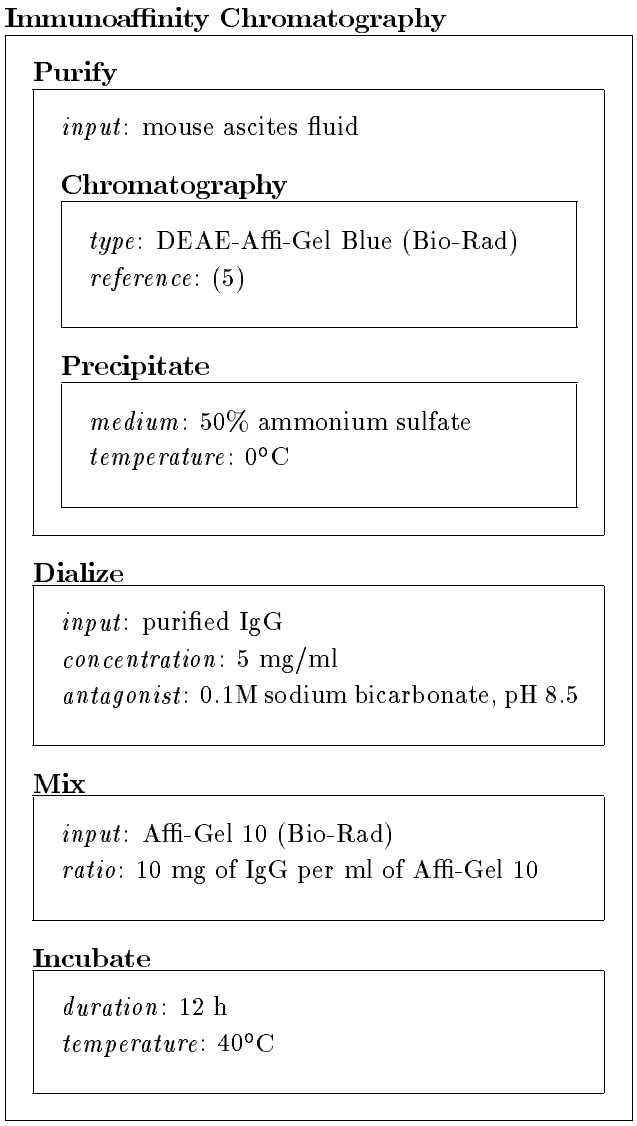

Figure 3: An example of a frame representation. The process schema shown in Fig. 2 describes the basic building block for the more complex set of processes shown here. The emphasis in this figure is on the *elaboration* relationship. The single **Immunoaffinity Chromatography** process is divided into four subprocesses, the first of which, **Purify**, is further divided into two subprocesses. The example here is a schematic, because the full representation involves much more detail as well as references to complex schema objects, not just words and phrases of the sort shown here.

```
<U.IMCR><SS1><ST>Immunoaffinity chromatography.</ST>
<P><U.S><U.PRFY>IgG was purified from mouse ascites fluid by
<U.CHRM>DEAE&ndash;Affi-Gel Blue (Bio-Rad) chromatography <RB>5</RB>
</U.CHRM> followed by <U.PRCP> precipitation in 50&percnt;
ammonium sulfate at 0&deg;C.
</U.PRCP></U.PRFY></U.S>
<U.S><U.DLYZ>Purified IgG (5 mg&divide;ml) was dialyzed against
0.1 M sodium bicarbonate, pH 8.5, </U.DLYZ>
<U.MIXD>mixed with Affi-Gel 10 (Bio-Rad) at a
ratio of 10 mg of IgG per ml of Affi-Gel 10, </U.MIXD> and
<U.INCB>incubated for 12 h at 4&deg;C.
</U.INCB></U.S></P></SS1></U.IMCR>
```

Figure 4: SGML representation of our sample text according to our schema. The text corresponding to a given object is bracketed, e.g., the Purify step text lies between the SGML tags, "<U.PRFY>" and "</U.PRFY>." The steps in this example text are nested inside each other. In general they will not be, because more than one text region may contribute to the information in a single object.

Application to Laboratory Notebook Software.

There are many possible applications for the tools we are creating. The one we have already talked about is the database to store information from Materials & Methods sections. Another important prospective use is to use the tool as a Laboratory Notebook. In other words, we should develop tools that assist a scientist in writing and storing Materials & Methods lab notebooks. Such notebooks should have information such as supplier, price and time required as well as the kind of information appearing in the published paper. Such a notebook could be used for the following:

1. Tracking use of materials and equipment by the laboratory.
2. Preparing the Materials & Methods section of their papers. In particular, this would allow reuse of material from one paper in another.
3. Serving as a means to identify and collect information about Materials & Methods from world-wide databases as well as to contribute information to them.

We are also developing tools for the generation of English text from knowledge frames. Of course, the generated text will be different from the original text. One can modify the generated text using standard word processing tools to conform to stylistic preferences.

Conclusions and Future Work.

The Materials & Methods schema we have discussed in this paper is not going to be static. It is going to evolve with the appearance of new papers. Hence, one of the big issues we are considering in this project is schema evolution. One should be able to add new types of objects to the schema and new relationships between types. As we have shown earlier, there is a category of objects corresponding to each process. So, when a new process is described in a paper, a new category needs to be added to the schema. There are also other examples when the schema evolution is necessary.

References

Andersen, P.; Hayes, P.; Huettner, A.; Schmandt, L.; Nirenburg, I.; and Weinstein, S. 1992. Automatic extraction of facts from press releases to generate news stories. In *Third Conf. on Applied Natural Language Processing*. Assoc. Computational Linguistics. 170–177.

Baclawski, K.; Mark, T.; Newby, R.; and Ramachandran, R. 1989. The nu& object-oriented semantic data modeling tool: preliminary report. Technical Report NU-CCS-90-17, Northeastern University, College of Computer Science.

Bairoch, A. and Boeckmann, B. 1991. The SWISS-PROT protein sequence data bank. *Nucleic Acids research* 19, Supplement:2247–2249.

Barker, W.C.; George, D.G.; Hunt, L.T.; and Garavelli, J.S. 1991. The PIR protein sequence database. *Nucleic Acids research* 19, Supplement:2231–2236.

Burks, C.; Cinkosky, M.J.; Fischer, W.M.; Gilna, P.; Hayden, J.E.; Keen, G.M.; Kelly, M.; Kristofferson, D.; and Lawrence, J. 1992. GenBank. *Nucleic Acids research* 20, Supplement:2065–2069.

Cinkosky, M.J.; Fickett, J.W.; Gilna, P.; and Burks, C. 1991. Electronic data publishing and GenBank. *Science* 1273–1277.

Finin, T.; McEntire, R.; Weir, C.; and Silk, B. 1991. A three-tiered approach to natural language text retrieval. In *Workshop Notes from the 9th Nat. Conf. on Artificial Intelligence. Natural Language Text Retrieval.* Amer. Assoc. Artifical Intelligence.

Higgins, D.G.; Fuchs, R.; Stoehr, P.J.; and Cameron, G.N. 1992. The EMBL data library. *Nucleic Acids research* 20, Supplement:2071–2074.

Jacobs, P. and Rau, L. 1990. SCISOR: Extracting information from on-line news. *Comm. ACM* 33:88–97.

Lehnert, W. and Sundheim, B. 1991. A perfomance evaluation of text-analysis technologies. *AI Magazine* 12(3):81–94.

McCartney, R.; Moreland, B.; and Pukinskis, M. 1992. Case acquisition from plain text: reading recipes from a cookbook. Technical Report TR-CSE-92-20, University of Connecticut Department of Computer Science and Engineering.

Pearson, P.L.; Matheson, N.W.; Flescher, D.C.; and Robbins, R.J. 1992. The GDB human genome data base Anno 1992. *Nucleic Acids research* 20, Supplement:2201–2206.

Stock, A. and Stock, J. 1987. Purification and characterization of the CheZ protein of bacterial chemotaxis. *Journal of Bacteriology.*