Project Semantic Web for Pathology
Report AP4

FU/NBI, Charite

May 17, 2005

Abstract

This report presents our achievements during the forth work package of the project "Semantic Web for Pathology", whose principal aim was to design and implement three application components, which present the semantically organized pathology information repository, consisting of patient reports and digital histological slides, to the user. The domain expert is able to query the archive using an ontology-controlled vocabulary. The results of his query are linked automatically to documents referencing related concepts in the ontology. Besides, for every pathology report the statistics component computes pre-defined reports which assist the diagnostic decision process, while the quality-assurance component evaluates its quality w.r.t. pre-defined guidelines.

Chapter 1

Introduction: Main goals of AP4

The AP4 is intended to realize the components which are really used by the pathologists in interacting with the pathology report archive. The corresponding documents have been annotated using ontological concepts by the transformation component (AP3) and can be therefore re-used with high quality and efficiency for further purposes, such as retrieval, quality assurance or statistical evaluations. The main application component in the emerging pathology information system is the retrieval component, which employs the underlying domain knowledge for various purposes in order to allow advanced search and presentation features with go beyond common keyword-based image and text retrieval techniques.

This report gives an account of our work towards the realization of these goals. Chapter 2 describes the functionality of the retrieval component. The statistic tool in presented in Chapter 3, while Chapter 4 accounts for preliminary considerations towards a Semantic Web based quality assurance module. Details about the implementation are available as JavaDoc at:

`http://www.inf.fu-berlin.de/inst/ag-nbi/research/swpatho/deutsch/javadocs.htm`

Chapter 2

The Retrieval Component

The main goal of our pathology information system was to reorganize the pathology information archive consisting of textual pathology reports and associated digital histological slides so that the valuable domain knowledge they represent implicitly could be optimally reused and shared in a variety of contexts. The main use scenario for the semantically annotated information items is retrieval. By means of the developed ontologies the system should be able to guide the user in expressing her queries, to rewrite user-defined queries for the optimization of the results and to offer semantically enriched methods to present and correlate these results.

A retrieval scenario consists of the following steps:

1. define search query: the user provides the search terms by using the ontology or as free text. Search terms are connected by common logical operators (AND, OR).

2. return results: the results are returned as a list containing the corresponding case identifier and a fragment of the text containing the search concept(s). Additionally a list of related terms (w.r.t. the ontology) and the number of relevant documents is presented to the user in graphical form. Given a specific documents, the user is provided statistical information, a summary of the quality criteria and the images corresponding to the document.

3. continue search: the user is able to select further relevant terms in the ontology as new search queries.

The ontology plays an essential role during the retrieval:

- it is used as query vocabulary: the user can select query terms directly from the ontology, which is visualized in tree form. Due to the size of the ontology, a pre-selection of a fragment of the ontology can be performed by the user, which may input a start concept. According to this information, a string-based comparison is executed in order to provide similar

concepts for the specification of the query. The string comparison is based on RDQL, a common query language for RDF-based data. A more sophisticated string comparison could be realized if more advanced query languages for RDF and OWL would be available and supported by Jena.

- it is used to refine the queries: the search is completed with further search terms corresponding to more general and more specific concepts (sub- and super-classes in the ontology). A refinement on the basis of pre-defined semantic relationships could also be integrated to the system with minimal effort.

- it is used to visualize the query results in the context of their usage. For every query concept the user is presented a sub-graph of the ontology which is annotated with the number of hits corresponding to each similar concept. By these means the user is provided an intuitive way to continue her search

- it is used to rank the hit list according to the number and importance of the contained search terms

- it is used to compute a similarity measure between documents (see below)

Chapter 3

The Statistics Component

Statistical data is an important part of medical research and helps to classify the individual diagnosis of a case in relation to the daily routine. The main scope of our statistical component is to build descriptive statistical reports for specific diseases/diagnosis based on the following criteria: age, gender and disease/diagnosis. Possible questions are e.g. the age distribution of a special disease or the most frequently diseases for a given age or gender. The other focus is to investigate the incidence of special findings in relation to a given diagnosis. For example asking about "How often occurs the diagnosis lung cancer in relation with vascular invasion?" is a possible scenario.
Our statistics tool is integrated in the report editor and the search tool. The statistical results are presented graphically as bar charts and in summarized text form (see Figure 3.1. The statistics component uses ontological knowledge in order to identify the diagnoses and diseases for each patient report, and patient data, which is already available in XML-HL7 form and stored in the report database.

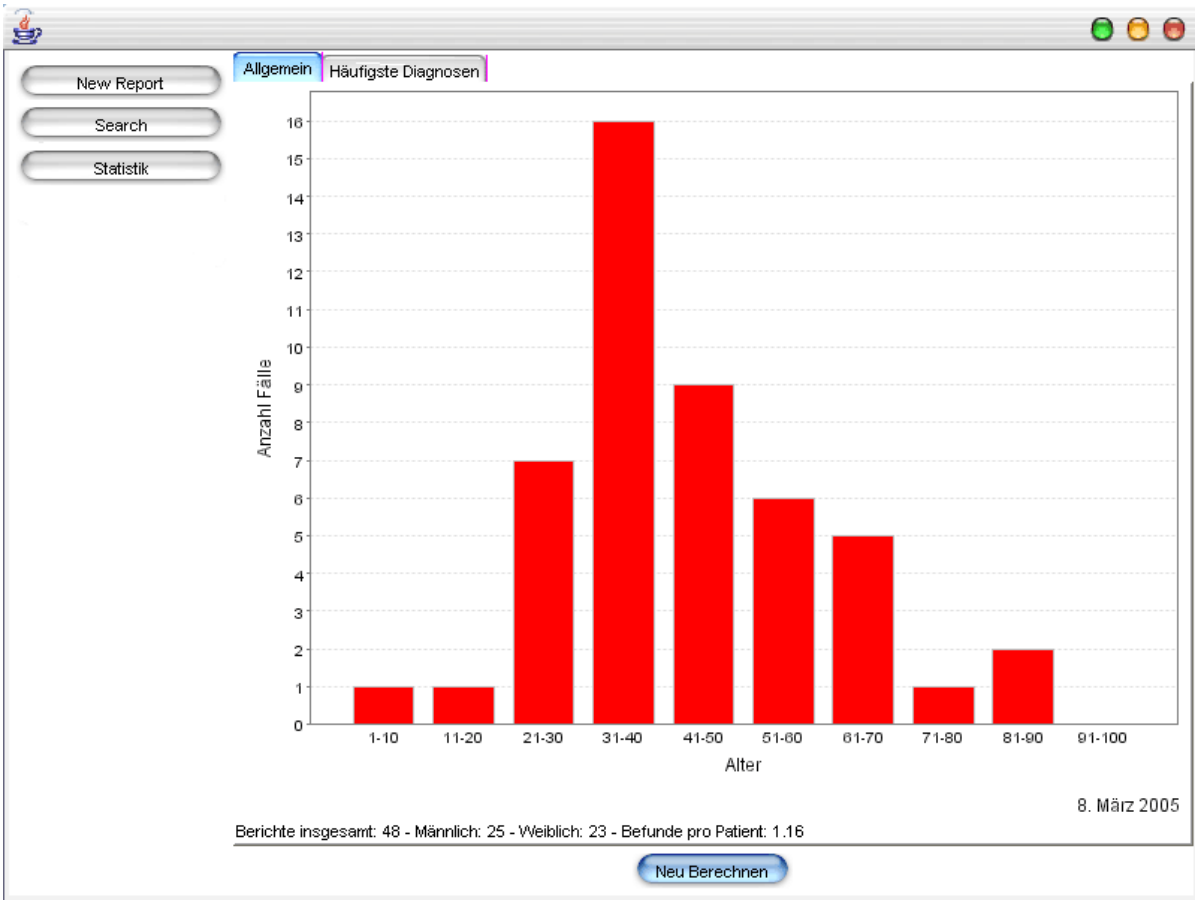

Figure 3.1: The Statistics Component

Chapter 4

The Quality Assurance Component

The quality of medical reports and in particular pathological reports is important for a frictionless inter- and intra-organizational information transfer. Furthermore the quality requirements to the pathological reports are especially high due to the following reasons:

- these reports are the basis for clinical treatment,
- the diagnosis and classification mentioned in the reports are included in treatment oriented studies,
- they are delivered to cancer centers for statistics and research and
- they are part of the quality assurance in other medical fields.

A tool for the quality assurance of patient records is therefore a reasonable extension for our report system. After analyzing the quality guidelines at the Charitè Institute of Pathology, we decided to define two levels for the quality tests:

1. Level1 relates to the general structure of the report (syntactical quality) and
2. Level2 concentrates the content of the report parts itself (semantical quality)

The result of the quality tests is presented as a check list (see Figure 4.1). The individual tests items are weighted by there importance. The cumulative quality test result is calculated as the sum of the weighted test results.

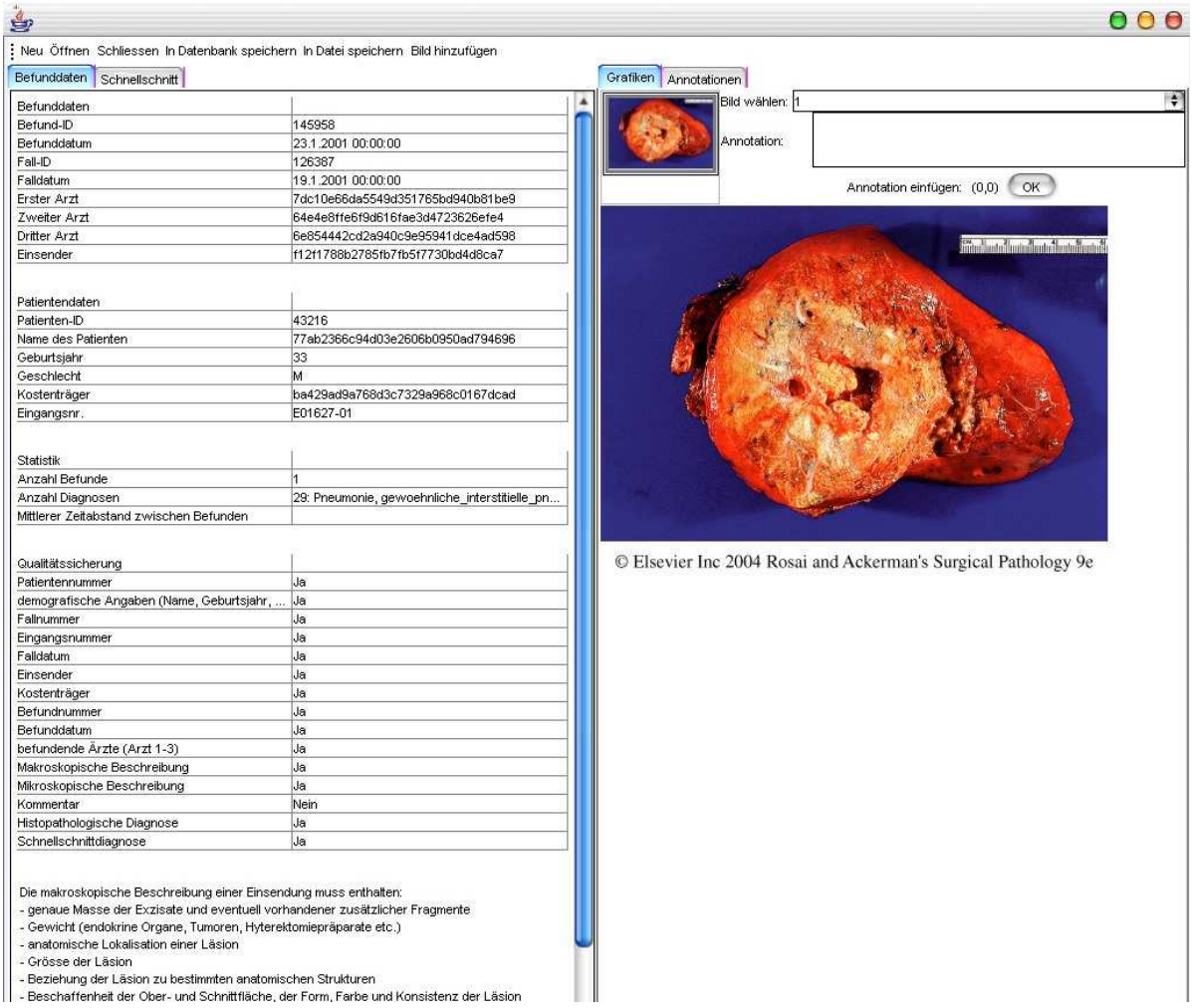

Figure 4.1: The Quality Assurance Component

4.1 Quality Assurance Level1

In this level the completeness of the administrative data and the presence of compulsory and optional report parts are tested. Examples of mandatory administrative data are patient id, case number, e-number (entry number), the sender etc. The required report parts are the macroscopic description and the histopathologic diagnosis section. The microscopic description is desired but not required.
From an implementation point of view the first level of the quality assurance is realized by checking the corresponding quality criteria directly on the patient records stored in XML-HL7 form in a Xindice database. For this level no ontological knowledge is taken into consideration.

4.2 Quality Assurance Level2

The second level of the quality assurance focuses on the content of the macroscopic description and the diagnosis/commentary section. Due to the optional character of the microscopic section we decided to leave it out for our prototypical implementation. The quality assurance tool tests the presence of the following information in the macroscopic description:

- the kind and the exact weight and volume of the pathologic compound,
- the anatomical localization of the lesion,
- the dimension of the lesion,
- the location of the lesion in relation to other anatomical structures,
- the appearance and consistence of the cutting area,
- the relation of the compound to the resection borders of the lesion.

The selection of the appropriate criteria to be included in the quality assurance procedure depends on the type of pathologic compound that is currently examined in the report (e.g. the fluid of a puncture does not have a dimension and cutting area).
In the diagnosis/commentary section the most important information is the diagnosis. The quality assurance tool tests the presence of a diagnosis and the answer to a possible clinical question in the macroscopy section. In case of particular diseases (e.g. lung cancer) a clinical classification is required and its presence will also be tested. For this purpose the domain experts examined a relevant subset of the report corpus to identify typical encodings used by pathologists to write down the ICD10 or TNM classification code of a certain diagnosis (see Appendix .1).

According to the examination the disease classification is encoded using the following pattern: $pT?, pN?, pM?, G?, R?, L?, V?$. The six parameters are as follows:

- pT is the tumor state,
- pN related to the lymph node state,
- pM relates to the metastasis state of the tumor,
- G is the tumor degree,
- R gives information about the resection procedure,
- L states for the microscopical infection of lymphatic vessels, and
- V states for the microscopical infection of the blood vessels.

A correct encoding requires the specification of at least one of the following parameters: pT, pN, G, R. Reports from 2004 and later have to additionally mention the type and localization of the tumor. The former is a combination of 5 digits separated by slash (e.g. "8070/3", while the latter starts with the capital letter "C" followed by a 3 digits separated by a colon (e.g. "C18.3").

In the commentary section we look for information about additional examinations. If mentioned, the presence of a second report has to be tested. Likewise disease classification the domain experts generated a list of typical expressions used to signal the presence of further examinations and reports. These expressions are used as input by a linguistic component which tests whether the corresponding quality feature is fulfilled by a given pathology report (see Appendix .2).
The realization of the second level of the quality assurance is a non-trivial task, because of the high requirements it imposes for the granularity and the domain coverage of the conceptual model and due to the ambiguity and subjectivity of certain quality criteria. While criteria related to the anatomical localization, the dimensions and the weight of the compound can be easily tested by simply querying the semantic representation of the reports, the appearance and the consistence of the cutting area as well as the issue of the resection borders require a fine-grained model of general-purpose spacial relationships (e.g. a solid corpus has borders and surfaces), which is not available in our ontology library at the desired level of precision. Extending the upper-level/generic part of our ontology library is subject of future work: after a first user-driven system evaluation we will extend the space ontology according to the user feedback. A third level of the quality assurance should focus on selected diseases such as bronchitis and use rules to detect whether the report mentions and takes into account all the necessary clues which are pre-defined for the given diagnosis.

.1 Examples of Disease Encodings

- I. Mäßig differenziertes, nicht verhornendes Plattenepithelkarzinom des rechten Lungenoberlappens. Retentionspneumonie in Nachbarschaft des Tumors. Tumorfreie bedeckende Pleura. Tumorfreie Resektionslinien. 8 tumorfreie anhängende Hiluslymphknoten. II. 2 tumorfreie Lymphknoten der Station 4. III. 6 tumorfreie Lymphknoten der Station 10. IV. Tumorfreier Lymphknoten der Station 12. **Tumorklassifikation: G2, pT2, pN0**.

- Bezüglich der nicht freien chirurgischen Abtragungsebene wurde im Schnellschnitt bereits Stellung bezogen. **Tumorklassifikation: pT1, G2, R1**.

- I. Unteres Bilobektomiepräparat von rechts mit einem mäßig bis wenig differenzierten, peripheren Plattenepithelkarzinom der Lunge mit geringer adenoider Differenzierung (max. 5%). 7 tumorfreie lobäre Lymphknoten. II. 2 tumorfreie Lymphknoten, n.A. Station 12. **Tumorklassifikation: pT2, pN0 (0/9), G3, L0, V0, R0**.

- I. Oberlappenmanschettenresektat von rechts mit einem mäßig differenzierten, verhornendem, zentralsitzendem Plattenepithelkarzinom mit beginnender Arrosion von Gefäßen. Massive Schleimretention und partiell abszedierende floride und chronisch-granulierende Entzündung im übrigen Lungenparenchym. Ein tumorfreier lobärer Lymphknoten. II. Tumorfreie Rippe mit Zeichen einer Frakturheilung, n. A. 8. Rippe rechts. **Tumorklassifikation: pT2, L0, V0, R0**.

- In diesem Resektat 7 Lymphknotenmetastasen in 7 Lymphknoten. Unter Einbeziehung der Vorbefunde ergeben sich 15 Lymphknotenmetastasen in 16 Lymphknoten, nach klinischer Markierung lobär und hilär. **Tumorklassifikation: pT2, pN1, G3, R1**.

- **Tumorklassifikation nach WHO: pT2, pN0, pMX; G2, R0**.

- **Tumorklassifikation: G2, pT2, pN0**.

- Abschließende Tumorklassifikation: pT4, pN2 (5/13), G3. Auf Grund des Primärtumorstadiums pT4 mit Infiltration des Perikards (vgl. E 17592/01) muss leider von einer R1-Resektion ausgegangen w

.2 Examples of verbalizations to indicate further examinations

- Die bisher durchgeführt immunhistologische Untersuchung spricht gegen das Vorliegen eines Karzinoms. **Wir werden noch weitere Untersuchungen durchführen und ein zweites Mal berichten**. Nach klin. Angabe vordiagnostiziertes Adenokarzinom.

Die bisherigen Untersuchungen sprechen am ehesten für das Vorliegen eines malignen peripheren Nervenscheidentumors (MPNST). Es handelt sich jedoch offensichtlich um einen seltenen Tumor, **wir werden daher noch ein Konsil anfordern und dann ein weiteres Mal berichten**.

- **Immunhistologie zur Diagnosesicherung folgt**. TNM-Klassifikation erfolgt in gleichen Befund.

- Tumorklassifikation: pT4, pN2, G3. Der Primärtumorstatus ergibt sich aus der Tumorinfiltration im Bereich des Perikards (vgl. E 17592/01). **Wir werden noch immunhistologische Zusatzuntersuchungen durchführen**, zum einen zur Beurteilung der Differenzierung des Tumors und zum anderen zur Festlegung des Lymphknotenquotientens; **Zweitbericht folgt**.

- Hier im Kommentart des ersten Berichtes kein Hinweis auf einen Zweitbericht, jedoch in der Mikroskopie des Zweitberichtes: **Wunschgefmäß haben wir weitere immunhistochemische Präparate angefertigt**.

- Nach klin. Angabe vordiagnostiziertes Nierenzellkarzinom vom Ductus-Bellini-Typ links, Z.n. OP am 12.1.01 (pT3a, G3, N1). Jetzt 9 x 5 cm große Raumforderung zentral links. **Zur histogenetischen Einordnung (insbesondere Differentialdiagnose pulmonale Metastase des Nierenzellkarzinoms verusus primäres Bronchialkarzinom) wird das Material noch immunhistologisch untersucht. Ein 2. Bericht folgt**. Kein kleinzelliges Karzinom.

- Tumorklassifikation: pT1, pN2 (4/5), G3, R0. Das histologische Bild ist verdächtig auf das Vorliegen einer neuroendokrinen Differenzierung. **Wir werden noch immunhistologische Zusatzuntersuchungen durchführen und dann ein zweites Mal berichten**.

- Aufgrund der adenoiden Differenzierung beider Karzinome spricht der Befund für eine unabhängige Karzinommanifestation in bezug auf das klinisch angegebene Hypopharynxkarzinom (Z. n. Hypopharynxkarzinom links pT3, pN1, M0, G3 mit Larynektomie und radikaler Neckdissection links 12/98). Da beide Adenokarzinome ebenfalls eine unterschiedliche Differenzierung aufweisen, handelt es sich bei dem Lungenkarzinom auch eher um unabhängige Tumormanifestationen. **Wir werden das Präparat noch weiter aufarbeiten und dann ein 2. Mal berichten**.

- Klinisch myelodysplastisches Syndrom, Verdacht auf ALL, bekannte Sarkoidose seit 4 Jahren, Verdacht auf Mykose (Aspergillen). Histomorphologisch kein Hinweis auf Malignität, das typische Bild einer Sarkoidose liegt nicht vor. Eine Pilzinfektion läßt sich in vorliegenden Schnitten und Färbungen nicht nachweisen, **wir werden das Material jedoch noch weiter aufarbeiten und dann ein zweites Mal berichten**.

  In den konventionellen Zusatzfärbungen kein Nachweis von Erregern, insgesamt läßt der Befund jedoch an eine infektiöse Genese denken. Kein

Hinweis auf die angegebene Sarkoidose. Fokal Zellen mit fraglichen Einschlußkörperchen, vereinbar mit CMV. **Wir werden noch immunhistologische und molekularbiologische Zusatzuntersuchungen durchführen und dann ein 3. Mal berichten**.

- Anteile eines malignen Tumors finden sich nicht. **Ergänzend immunhistologische Untersuchungen, wir berichten erneut**.

- **Immunhistologische Untersuchungen folgen, wir berichten erneut**.

- Diskussion des Befundes vorab am 01.11.01 um 16.00 Uhr, Dr. XXXXX. Für malignes Tumorwachstum ergibt sich kein Anhalt. **Weitere Untersuchungen werden noch durchgeführt. Wir berichten erneut**.

- Die bisher durchgeführt immunhistologische Untersuchung spricht gegen das Vorliegen eines Karzinoms. **Wir werden noch weitere Untersuchungen durchführen und ein zweites Mal berichten**. Nach klin. Angabe vordiagnostiziertes Adenokarzinom.

  Die bisherigen Untersuchungen sprechen am ehesten für das Vorliegen eines malignen peripheren Nervenscheidentumors (MPNST). Es handelt sich jedoch offensichtlich um einen seltenen Tumor, **wir werden daher noch ein Konsil anfordern und dann ein weiteres Mal berichten**.

  Der Tumor ist wahrscheinlich lokal im Gesunden reseziert, entsprechend einer R0-Resektion. **Die endgültige Klassifizierung des Tumors erfolgt nach Vorlage des externen Konsils**. Mit großer Wahrscheinlichkeit handelt es sich um ein Sarkom. Die von uns an den Lymphknoten (E-Nr. ???? und ????) sowie der Zytologie (E-Nr. 38332/01) gestellten Diagnose eines Karzinoms ist daher vermutlich zu revidieren. Telefonische Befundbesprechung mit Fr. Dr. XXXX von der Klinik YYYY am 31.10.01.

- **Zur Bestätigung der Diagnose und näheren Typisierung des Lymphoms werden wir noch weitere immunhistologische und molekulargenetische Zusatzuntersuchungen durchführen und dann ein 2. Mal berichten. Weiterer Zusatzbericht nach molekulargenetischer Zusatzuntersuchung**.