Towards Automatic Identification of Completeness and Consistency in Digital Dossiers

Martijn Warnier Frances Brazier
VU University Amsterdam
Intelligent Interactive Distributed Systems
De Boelelaan 1081a
Amsterdam, The Netherlands
{warnier,frances}@cs.vu.nl

Martin Apistola Anja Oskamp
VU University Amsterdam
Computer/Law Institute
De Boelelaan 1105
Amsterdam, The Netherlands
{m.apistola,a.oskamp}@rechten.vu.nl

ABSTRACT

The emergence of digital dossiers in Courts of Law presents new opportunities to streamline the criminal prosecution chain. This papers proposes the use of agent technology to support automatic verification of consistency and completeness of data in such dossiers. It sketches how agent systems in combination with other AI technology, can be used to enforce consistency and completeness in digital dossiers in the context of the semi-open environment of the Courts.

Categories and Subject Descriptors

H.2.4 [**Information Systems**]: Database Management—*Distributed databases*

1. INTRODUCTION

Technology is changing today's law practice. The use of digital dossiers by the Public Prosecution and Courts of Law in trials is an example. Consistency and completeness of digital dossiers are major challenges in this context. In the semi-open environment of the Courts these challenges are magnified: different sources of information distributed both physically and across organizations need to work together within fixed boundaries set by the law to compile a dossier. Potentially good solutions to ensure consistency and completeness of digital dossiers may not be sufficient if the characteristics of the semi-open environment are not taken into account. This paper identifies important issues with regard to the completeness of digital dossiers and consistency of data. Possible solutions for automatic completeness and consistency checking of digital dossiers, based on AI techniques, are discussed. Agent technology, one of paradigms deployed within AI, is used to efficiently apply these techniques to the digital dossier.

The digital dossier is the focus of the Agent-based Criminal Court Electronic Support Systems (ACCESS) project[1]. The ACCESS project aims to explore the feasibility of a fully distributed system, to support users based on knowledge of user preferences, to support information sharing between users, and to improve the overall efficiency of current practice. Deploying agent technology to ensure completeness and consistency of digital dossiers using data scattered over a large number of different physically distributed organizations is one of the main objectives of the project. Agent technology [5, 8, 17] is a promising and enabling technology in such large scale distributed environments. Artificial Intelligence, for instance, when embedded in agent technology, offers new possibilities with regard to the digital dossier.

The remainder of this paper is organized as follows. Section 2 provides a motivating example that illustrates the complexities involved in a criminal prosecution chain for a juvenile repeat offender. Section 3 introduces the digital dossier and describes the context of the semi-open environment as holds for Courts of Law. Section 4 then explains the main issues and possible solutions for consistency and completeness in the semi-open environment of the Courts. The paper ends with a discussion and conclusions.

[1] `http://www.iids.org/access`

ICAIL '07, June 4-8, 2007, Palo Alto, CA USA

2. THE CRIMINAL PROSECUTION CHAIN OF A JUVENILE REPEAT OFFENDER

A trial and its preparation involves numerous organizations. This process starts once someone is suspected, continues through trial preparation and ends with a verdict and execution of a sentence. This whole process is called a *criminal prosecution chain*. An example of a criminal prosecution chain for a juvenile repeat offender is illustrated below. This example has been constructed from an actual case. All personal information has been anonymized and numerous details have been omitted. However it does illustrate the complexities involved in such cases. Note that this scenario describes the current Dutch situation, legal constraints are also analyzed in a Dutch legal setting, according to Dutch law.

This criminal prosecution chain starts when the Police arrest a juvenile suspect for vandalism. The subject is escorted to the Police station where an assistant prosecutor questions the suspect. The Police open a new dossier specifically for this case. This dossier contains a summary of the offense for which the subject is being charged, the date and location of the incident, number of suspects, personal data of the victim, the official police report, and other relevant information. The suspect then becomes the subject of investigation: the personal data he/she has provided is cross

referenced with the Municipal Database.[2] The Police also queries local Repeat Offender Databases to discover whether this subject is a known repeat offender. As this case concerns a minor, a request is issued to other organizations for juvenile offenders, to provide relevant information about the minor's background. All of this information is added to the dossier.

After collecting this information, the Police and the Assistant Prosecutor inform the Public Prosecutor of the case and transfer the dossier. The Public Prosecutor decides whether to press charges or, to pursue an alternative if other (minor) punishment is deemed more suitable. This decision is based both on the specific details of the current case and the (criminal) history of the offender. A dedicated Judicial Documentation Database is used to retrieve information on the criminal past of the subject. Typically, at this point, the Public Prosecutor will again consult Municipal Databases and local Juvenile Repeat Offender Systems. All information is cross referenced with the case dossier and information is updated when needed.

If the Public Prosecutor decides to bring the case to court, as is the case, the next mandatory step involves informing the Council for Child Defence. In the Dutch context the Council for Child Defence has the task to investigate all crimes of minors. In addition to the criminal offenses of the minor, the family situation and other relevant social factors are taken into account. This results in a motivated advice for suitable punishment of the subject. This advice is added to the dossier. The prosecutor then serves a summons and a lawyer is assigned to the juvenile subject. Adding the summons to the dossier finalizes the dossier at this point. A copy of the dossier is sent to the Court and to the lawyer of the subject.

To check the correctness of the dossier, the presiding judge may query judicial history and other judicial documentation in the Judicial Documentation Database as well as information from Municipal and other databases.

At the court session the information in the dossier is used by all parties involved. The Public Prosecutor demands a suitable sentence, the lawyer presents the defense and ultimately the judge comes to a verdict. The subject is sentenced and all information regarding the court session is added to the dossier. The dossier itself is filed in Judicial Documentation Database for future reference.

3. THE DIGITAL DOSSIER

The example scenario presented in the previous section illustrates that not only numerous parties are involved, but also many databases are consulted in a rather simple case such as that of a juvenile repeat offender, often more than once. The result is often a sizable dossier. The Courts of Amsterdam and Rotterdam are currently exploring the potential of a digital version of this dossier in a pilot study. The digital dossier can be seen as the center of a large distributed system that consist of all organizations, such as Police, Courts, municipal databases etc., see Figure 1.

The information comes from many different sources: the combined systems of the Courts, the Public Prosecution, the Police and all other parties together form a *semi-open*

[2]This database contains, for each citizen of a municipal, personal information such as name, date of birth, sex, marital status, and family relationships.

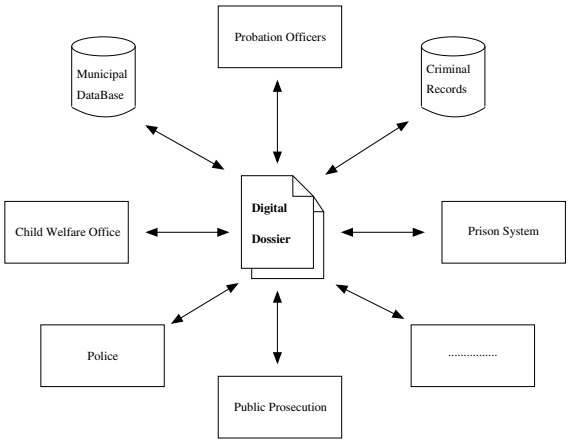

Figure 1: The Digital Dossier depends on –and interacts with– data from a large number of distributed databases.

system, not quite an open system, nor a completely closed system. An *open system* is a computer system that is configured to allow access to outside parties. In contrast, in a *closed system* only known users are allowed access (after authorization) to parts of a computer system. The systems used to compile a digital dossier, are neither closed nor open: the organizations involved are known, but are not allowed to access each others' systems/databases. Each individual organization is responsible for its own systems, and information provision. Requests are honored, systems trusted. Sometimes information from outside sources is used.

The Police, the Public Prosecution Service, the Courts and the Prisons together form the backbone of the Dutch criminal prosecution chain. As illustrated in the example in Section 2 there is an order in this chain. The Police and the Public Prosecutor are responsible for the criminal investigation. The Public Prosecutor is responsible for the prosecution followed by the Court. The Court is responsible for judging cases. The Prisons are responsible for executing sentences. (Note that the latter organization does not play a role in the above scenario). Each organization in this chain depends on the information provided by organizations consulted in previous phases of the criminal prosecution chain. These include organizations such as Municipalities and the Counsel for Child Defense. Each organization involved in the criminal prosecution chain uses its own computer system/databases to process, store and share information. Currently digital information exchange between organizations is being explored/pursued.

Each organization's computer systems can be seen as stand alone systems, i.e., as closed systems. Only police officers can access the computer system used by the Police and only employees of the Public Prosecution can access (parts of) dossiers at the Public Prosecution. However, as all of these systems (can) exchange information, the chain of systems is no longer a closed system. The characteristics of each site can vary considerably: they can have different procedures for access, reliability and/or security. For example, a Municipal Database that contains name and address information has other goals (and thus system characteristic) than a computer system that stores DNA of individuals. Moreover, at

several places in the system there are outgoing connections to public (open) computer systems, e.g. systems that provide up to date information concerning the current state of a law etc. Thus the system as a whole constitutes an example of a semi-open system: a system that is not completely closed, but certainly also not completely open.

Semi-open systems are challenging for numerous reasons. It is complex to ensure in such systems that privacy sensitive data is well protected against malicious intent, and to ensure that other less sensitive information is available to a more general public [13]. Such systems generally tend to make *all* data harder to access, including the less sensitive information that can be of interest to a large public. This phenomenon is known as 'label creep' in literature [11]. Other security related problems of semi-open systems deal with access control, confidentiality and integrity of data, see [16] for a discussion of the security requirements in such a system. From a legal perspective, it not clear what kind of integration of the individual systems is allowed by (Dutch) law.

For this reason the digital dossier is currently based on paper exchange/fax copies of files. Currently, for this pilot, (paper) versions of all files are scanned and stored as pdf-files in the digital dossier. A web-based user interface allows a user to access the digital dossier, take notes by marking up their own copy of files, and share notes if so wished. A future version of the digital dossier will also incorporate machine parsable (XML like) content, including multi-media material (sound, images and video).

The ACCESS project assumes the above mentioned problems will be solved, and that digital information exchange will be possible. The result is thus an example of a distributed system: a system with physically distributed sources, distributed across organizational boundaries. A distributed digital dossier [16] is then an option. Information is managed by the authority responsible for the data. A new instance of a digital dossier is created by the Public Prosecutor when a new case is presented and the Public Prosecutor decides to prosecute a defendant.

A newly created dossier consists of records and meta data. The meta data contains information such as the access control list (who may read and alter the dossier), what type of information needs to be present in a dossier (for completeness checks, see Section 4). The meta data part of the dossier is stored in the database at the Public Prosecution.

Individual records are distributed over the responsible organizations: personal information is maintained by the Municipal Database, family related information for juveniles is maintained by the Council for Child Defence etc. This ensures that information in the digital dossier is kept as up-to-date as the information known to the responsible organizations. When information changes this is flagged by the responsible organization and synchronized with the dossier at the Public Prosecution.

Thus the dossier itself is stored at the Public Prosecution while relevant records are maintained and stored by the responsible organizations and then synchronized with the digital dossier at the Public Prosecution, when necessary. Consistency and completeness are, however, aspects that need to be guaranteed.

An agent-based system has been designed to support access to the distributed digital dossier, completeness and consistency. Rather then using 'classical' distributed information systems [10, 12, 15] agent technology has been chosen because it provides a means to distribute responsibilities and tasks across interacting distributed autonomous systems. Moreover, agent systems are inherently modular –since separate functionality can implemented by specific, dedicated agents– and are ideal for deployment in large scale distributed environments.

Dedicated software agents can be programmed to perform the task of guarding consistency both within a source and between sources within the digital dossier, and notifying appropriate (human) parties when needed. With respect to completeness dedicated agents can monitor the availability of necessary documents in the digital dossier. For instance, a trial cannot start if a copy of the original police report is not in the digital dossier. Software agents can guard such completeness issues [14].

4. COMPLETENESS AND CONSISTENCY

An information management system as used by the Courts needs to adhere to a large range of requirements of which the most important are: completeness, consistency, security and reliability . The focus of this paper is on completeness and consistency. In [16] the security issues related to the distributed digital dossier are discussed. Reliability can in general be guaranteed by introducing enough redundancy in the system, such as local (secure) caching and multiple (physical) network connections between sites. The remainder of this section sketches how AI techniques can be deployed to enforce completeness and consistency checks. The emphasis is not be on the particular details of the individual techniques, but rather on how multi agent systems can be used to combine techniques efficiently.

Completeness requirements can be formulated for each type of offense. Some information, such as personal information regarding the defendant as well as the offense for which the defendant is charged and the original police report need to be included in each and every dossier. In addition, supplementary offense specific information is often required, e.g. as the scenario in Section 2 illustrates, in a case involving a minor the Council for Child Defence needs to investigate the minor's past and his or her family/social situation.

Automatic completeness checks of dossiers involves the following:

- Determine for each *type of offense* which information is mandatory.
- Check if all the mandatory information is in the dossier before it is sent to the Court and the defendant's lawyer.

Determining which information is mandatory for each type of offense is a mainly static process, although laws can change. As a result, this information may need to be updated during the course of an investigation. A domain expert needs to identify which information is mandatory. Knowledge acquisition [1] techniques can be used to support this process. The domain of, for example, all legal cases involving juveniles, is very large. Therefore some kind of automated techniques, e.g. based on clustering [6], can be used for an initial categorization of mandatory information in dossiers. This categorization should ultimately form a legal ontology [7]. This ontology is then stored in the meta-data of a dossier, allowing software agents to check completeness.

Automatically checking the completeness of a dossier is a more dynamic task. A dedicated software agent is ideal for this. Each dossier can have its own agent responsible for checking completeness, that can warn a human user that a dossier is not complete. An incomplete case can not be sent to the Court.

Internal consistency of data[3] in individual dossiers forms an important requirement which should be possible to enforce when the data is available digitally. Thus, if an organization, e.g. the Council for Child Defence, adds a record to a dossier, the system checks whether the information is consistent with all other information in the dossier, using e.g. personal information, such as name, address, age and sex of the subject. Consistency also entails checks for reasonable entries in data fields, e.g. the age of subjects should be in the range 0–120, and other checks, e.g. some bank account numbers and (Dutch) social security numbers allow simple tests (comparable to checksums) that can very if the number represents a possible value.

Checking internal consistency is, in fact, strongly related to the *matching problem*. Simply checking to see if e.g. new personal data added to a digital dossier matches existing data is not enough. Even if a dossier is well structured, and all system interaction is based on the same common ontology, a large information management system [4], may not discover inconsistencies such as those due to mismatches caused e.g. by typing errors. If a system does not recognise the similarity between e.g. the name in a digital dossier and the name submitted, the receiving system may, for example, create a new entry in a database. This problem needs to be considered, especially given the semi-open environment of the Courts. A more robust method is needed to enforce internal consistency in digital dossiers: *fuzzy matching techniques*[2] provide a means to identify partial matches. It is not clear if automated systems can be based on these results.

Enforcing consistency is a far more dynamic problem then completeness checking. Especially in the case of a distributed digital dossier where data in a record can change at any moment before the dossier is sent to the Court. Again, agents can be used to enforce consistency checks. A dedicated agent can be made responsible for the consistency checks of each individual dossier, interacting with dedicated agents responsible for the information at the source. This central agent should only allow the addition and/or altering of information if this is consistent with the remainder of the dossier. When conflicts arise a human user needs to be notified of this and further handle the inconsistency.

Checking consistency between several dossiers, e.g. in cases with multiple suspects, can be handled in a similar matter as described above. As long as there is a clear ontology, a dedicated agent can be assigned to the task of checking inter-consistencies of dossiers.

Consistency and completeness of data in large information systems is a notoriously hard problem. In the context of the Courts, inconsistencies and incompleteness of data can lead to postponing trials or even a mistrial. Thus, automatically identifying incompleteness and inconsistencies in dossiers can have large benefits for the (clogged) legal system. Automatically fixing (part of) these problems can have similar benefits.

[3]Internal consistency of a digital dossier should not be confused with data integrity of a digital dossier. The latter ensures that data is preserved during operations, e.g. when data is transferred over a network. While the former addresses the problem of preventing conflicting information in one dossier.

5. DISCUSSION AND CONCLUSIONS

The above section discusses a number of technical aspects associated with the completeness and consistency checks in the digital dossier. However, due to the specific domain of a Court of Law, some legal constraints also have to be taken into account.

The legal requirement of consistency and maintaining the integrity of the dossier becomes more urgent when dossiers are digitally built, stored and managed. To ensure that information is only added by an authorized entity, adding and changing information must be traceable. It must also be clear who has the right of access when, and what the rights of this party are. It is clear that access of the lawyer of the defendant is very limited, and usually restricted to the phase when the dossier is finalized[4]. His/her rights to make any changes will be even less, or virtually non-existent. Yet, the status of a document can depend on how much it is needed in a specific phase of the procedure. During custody, for instance, the defense may need direct access to hearings of witnesses.

With regard to the legal requirement of completeness the Dutch Supreme Court has formulated a criterion regarding documents in the dossier: to the extent that documents may influence evidence, it must be assumed that the Defense, the Judge and the Public Prosecutor will add all documents regarding the research to the dossier. All documents that may reasonably be relevant to the position of the suspect must be included in the dossier. The Judge can, when requested by the Defense and the Public Prosecutor, officially order certain documents to be added to the dossier. Data concerning the use of special research methods such as infiltration and (wire)taps, even if they do not lead to relevant results, must always be added to the dossier [3].

Another issue, that needs to be addressed in future research, concerns the legal implications of exchanging information between several organizations in the semi-open environment of the courts. At the moment it is not clear, from a legal perspective, what information (if any) may be shared automatically and what information can only be shared after an explicit approval of an authorized party.

Finally, an information management system for the Courts not only needs to address the issues concerning completeness and consistency of the digital dossier. Other requirements, in particular security requirements [16] but also e.g. reliability and user-friendliness of the resulting system, need to be considered. At the moment a proof-of-concept implementation of such a system is being developed using the AgentScape [9] agent platform.

Acknowledgments

This research is supported by the NLnet Foundation, `http://www.nlnet.nl`, and is conducted as part of the ACCESS project, `http://www.iids.org/access` funded by the NWO TOKEN program.

[4]A *finalized dossier* contains all information, chosen by the public prosecutor, that forms the basis for a trial. Information in such a dossier is no longer updated regularly, hence the name. See [16].

6. REFERENCES

[1] J. Boose. A survey of knowledge acquisition techniques and tools. *Knowledge Acquisition*, 1(1):3–37, 1989.

[2] S. Chaudhuri, K. Ganjam, V. Ganti, and R. Motwani. Robust and efficient fuzzy match for online data cleaning. In *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, pages 313–324, New York, NY, USA, 2003. ACM Press.

[3] G. Corstens. *Het Nederlands strafprocesrecht*. Kluwer, 2002.

[4] C. A. Ellis and G. J. Nutt. Office Information Systems and Computer Science. *ACM Comput. Surv.*, 12(1):27–60, 1980.

[5] J. Ferber. *Multi-Agent Systems: An Introduction to Distributed Artificial Intelligence*. Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA, 1999.

[6] D. Fisher. Knowledge acquisition via incremental conceptual clustering. *Machine Learning*, 2(2):139–172, 1987.

[7] A. Gangemi, M. Sagri, and D. Tiscornia. A Constructive Framework for Legal Ontologies. *Internal project report, EU 6FP METOKIS Project, Deliverable*, 2004.

[8] M. Luck, P. McBurney, and C. Preist. *Agent Technology: Enabling Next Generation Computing (A Roadmap for Agent Based Computing)*. AgentLink, 2003.

[9] B. Overeinder and F. Brazier. Scalable Middleware Environment for Agent-Based Internet Applications. In *Proceedings of the Workshop on State-of-the-Art in Scientific Computing (PARA'04)*, volume 3732 of *Lecture Notes in Computer Science*, pages 675–679, Copenhagen, Denmark, June 2004. Springer.

[10] M. Ozsu and P. Valduriez. *Principles of distributed database systems*. Prentice-Hall, Inc. Upper Saddle River, NJ, USA, 1991.

[11] A. Sabelfeld and A. C. Myers. Language-Based Information-Flow Security. *IEEE Journal on selected areas in communications*, 21(1), 2003.

[12] R. Sprague Jr and B. McNurlin. *Information systems management in practice*. Prentice-Hall, Inc. Upper Saddle River, NJ, USA, 1986.

[13] E. Stone and D. Stone. Privacy in organizations: Theoretical issues, research findings, and protection mechanisms. *Research in Personnel and Human Resources Management*, 8(3):349–411, 1990.

[14] L. Storchi. Het controleren van het digitale dossier op volledigheid met behulp van software agenten. Master's thesis, Vrije Universiteit Amsterdam, IIDS group, sept 2005.

[15] A. Tanenbaum and M. Van Steen. *Distributed Systems: Principles and Paradigms*. Prentice Hall PTR Upper Saddle River, NJ, USA, 2001.

[16] M. Warnier, F. Brazier, M. Apistola, and A. Oskamp. Secure Distributed Dossier Management in the Legal Domain. In *Proceedings of the 2nd International Workshop "Dependability and Security in e-Government" (DeSeGov 2007)*. IEEE, 2007.

[17] M. Wooldridge and N. Jennings. Intelligent Agents: Theory and Practice. *The Knowledge Engineering Review*, 10(2):115–152, 1995.