Digital Identity and Reputation in the Context of a Bounded Social Ecosystem

Ben Jennings and Anthony Finkelstein

University College London
London, UK
b.jennings@cs.ucl.ac.uk and a.finkelstein@cs.ucl.ac.uk

Abstract. This paper highlights the problem of digital identity, or cross-set unique identifying tokens, inherent in the application of social software in business processes. As social software, via blogs, wikis and other Web 2.0 software, starts to bring value to the enterprise, there is a need for a unified digital identity resource. From this basis, concepts of reputation and trust may be leveraged in the context of human agents working with such software and in larger workflow processes. By analysing human agent activity within existing data sets and providing a mechanism for adding new data, it is possible to correlate human agent activities and data creation via a digital identity across disparate sources of data. From this basis, new business processes may be created using this deeper understanding of human agents. This paper highlights the need for digital identity and presents a novel method for extracting digital identity patterns from heterogeneous data sets in an automated manner.

Key words: Identity, Reputation, Flexible, Workflow, BPEL, SOA, RMR, Recommendation, Trust

1 Introduction

The wide spread popularity of social software in the public technology space via the Web 2.0 meme has seen large scale user adoption, for example the Wikipedia project which has in excess of two million user submitted English peer reviewed articles [36]. In the enterprise context, the appeal of integrating social concepts is two fold: there is the opportunity to improve business processes through richer, socially enabled software interactions and to create mechanisms for the human agents within the enterprise to add value to the knowledge of the company. Extending the consumer facing meme, the Enterprise 2.0 term [20] is gaining adoption.

When integrating social software into business processes, there is the potential for a fundamentally deeper understanding of the individual within the enterprise. With the growing prevalence of Service Oriented Architecture (SOA) and web services based interactions, the integration with human agents is a problem of critical importance. The contemporary SOA solutions offer abstractions

to human agents in the form of worklists [7] which vary from a simple queue to a predefined grouping of related human agents. With the adoption of social software, via blogs, wikis, forums etc. there is a new source of data from which to mine information about specific individuals and aggregate that knowledge into groupings. By combining social software, data mining and natural language analysis, integrated business processes will have the potential to make a significantly more informed choice about the most appropriate human agent for a task.

A fundamental issue to overcome in this information rich enterprise environment is that of digital identity. In the context of this paper, the term digital identity is used to express the concept of a cross-set unique token. In any system, or set of integrated systems, used within a business process, establishing the exact identity of a human agent within that system, or subset thereof, is critical. When using social software data artefacts, having specific knowledge of who created that data, via a unique token or digital identity, allows a process to link a human agent to a specific body of work or expertise. If the enterprise were a green field environment, an integrated identity solution could be used. The more pragmatic approach in a real world scenario will be that of many legacy data artefacts (mailing lists, commit logs) with the addition of new, best of breed applications. For a unified view of the identity of human agents within such an enterprise, new techniques must be applied to analyse the existing data sets, and the data for potential pseudonyms and present a digital identity resource from which to make human agent based assertions.

This paper sets out to present an overview of issues associated with leveraging social software and identity within the enterprise. From this basis, the paper will suggest a novel mechanism from which to resolve digital identities to create a unified unique token to identify human agents across data sets. In the next section of this paper, digital identity will be put in context and an overview of the relationship between identity, reputation and trust will be presented. In the following section, social software and the integration of unified digital identity will be discussed. In the next section, the novel mechanism for addressing the problem of resolving digital identity in a legacy, heterogeneous and bounded ecosystem will be presented. In the final section, the application of this method in a larger context will be examined.

2 Digital Identity, Trust and Reputation

For an enterprise to integrate social software into business processes, there is the need for both reputation and an authoritative voice [6]. Without these tenets, there can be little value added to the enterprise. This section will present some issues when attempting to leverage social interaction as a facet of enterprise software. A major concern is that of information overload to which a specific human agent may be subjected and the continuing importance [16] of reputation, trust and digital identity in a bounded ecosystem.

2.1 Information Overload

As the adoption of social software within the enterprise domain becomes more prevalent, so comes the issue of too much information. As more human agents generate more blog posts, wiki edits and messages to mailing lists, this can lead to to information overload [18]. Studies have been performed that show if people are subjected to an overly dense information stream [25], they will be less productive, and in the worst case, abandon the system altogether. Any system hoping to provide information to human agents needs to consider this issue. Rather than modelling data connections based on existing physical systems, computer based mechanisms can act as a filtration system, distinguishing good information from less important.

Social software needs to provide contextually useful information connecting specific human agents to each other in order to respond to environmental information and subsequent business process exceptions. During the process of integrating social software (and existing social data artefacts such as mailing lists) within the enterprise, this overload issue may be mitigated by unifying identities within disparate silos of information in order to ascertain relationships between individuals and data. From this unified identity, less relevant data can be occluded from certain human agents, effectively bubbling up more relevant content and agents.

2.2 Digital Identity as a Facet of Reputation

Trust and reputation are subjective measures, as both are based entirely upon personal feelings and the interpretation of ambiguous signals [13], rather than the objective representation of fact. Trust may be viewed as a function of the agent's desire for an outcome in relation to their perception of the transactional risk dependant upon that agent's attitude towards risk in a specific context [23]. This measure may alleviate concerns of opportunistic behaviour [4] from the other participant in a given transaction. This opinion led abstraction of a deficit of information [27] can form the foundation of a decision making process. The act of aggregating individual human agent's interpretation of ambiguous or asymmetric knowledge [1] can lead to a broader context from which to make a decision.

Reputation is a socially constructed label [38] to try and present an interpretation of underlying intent. From an empirical perspective, reputation may be considered as an observation of past behaviour [33]. Reputation may also be viewed as an incentive to provide a positive transactional experience [35]. From the human agent judgements regarding vices and virtues, strengths and weaknesses [11] a subjective measure, either from an individual or aggregated viewpoint, provides a perceived quality of a specific characteristic [24]. As with all transactions predicated upon human agent interaction, there is the potential for manipulation via gaming the system. It is also important to note, when making reputation observations, that it would be incorrect to assume rational behaviour from all interacting agents [2]. Neither trust nor reputation express

factual or repeatable information, merely a subjective, individual or aggregated opinion of human agents within a specific context [14].

Without a clear sense of identity, there can be no foundation for trust or reputation. The value of trust within an online social environment has been the subject of much research within Epinions [17], Ebay [28] and movie recommendations [15]. In the enterprise environment, trust and reputation will also become a matter of key concern with the adoption of social software. Without a clear concept of identity across data artefact sets, there can be no concept of a unified human agent reputation. This concept of reputation, based on data generated through social software interaction, also raises an interesting question that is a continuing one for Ebay [9]. As reputation and trust are built on a continuing relationship between human agents and data, or physical items in the case of Ebay, if there are two human agents, one of whom has many low value interactions with an enterprise social system, and another agent who interacts infrequently but at high value, the high value agent will be more likely to be lost amongst the noise of the first agent. One technique to help alleviate that issue is by looking across data sets for transactions by the same person. By aggregating human agent interaction, the likelihood of divining higher value interactions increases.

In this section, the importance of digital identity as a fundamental building block of trust and reputation have been presented. The possibility for using an aggregated identity across heterogeneous legacy assets has been shown to act as an ameliorating factor for the adoption of social software. In the next section, some of the potential applications of social software in the enterprise will be discussed and a specific solution space predicated upon a uniform digital identity will be examined.

3 Social Software and Unified Identity

The application of social software to the enterprise should not be perceived as a silver bullet [5]. The act of adding a wiki or opening up a company's Subversion repository and accepting tickets will not, by default, leverage the social graph. In order to bring meaning to the integration of social software in an enterprise context, the system must strive to engage in conversation [22] with its users. From this basis, a business process can use such information about both the human agents and the data with which they interact to provide a cornerstone to integrate those agents into expert driven business processes. When the users of the social system perceive value from the interaction, more trust will be given to the system which will, in turn, encourage adoption. In order to put the necessity of a unified digital identity in context, the following section will look at some of the specific areas of potential application of social software and the potential enterprise benefits. At the end of the section, a specific domain usage of leveraging social software and digital identity will be discussed.

3.1 Leveraging Social Software Practices

Some of the early efforts in integrating social software into the enterprise have come from instant messaging [32] and blogging [21]. With the use of blogging in enterprise some patterns may be observed: observer, external creator and internal creator. There are obvious enterprise concerns over private company information, so two common paths that are taken are either limiting the creation and aggregation of content to behind the firewall (internal creator) or having strict policies [10] in place for outward facing content (external creator). The observer pattern can be seen as a human agent within the enterprise monitoring socially created content that relates to the business via aggregated keyword searching from a service such as Technorati. Another example of social software in the enterprise is the Dogear project. After lack of user adoption of a traditional fixed ontology system, IBM created a lightweight, social tagging system [26].

Referring back to the previous section of this paper, unless users adopt a system, the power of the social aspect will not be fully realised to the business processes. As has been seen, the adoption of social software tends to be conservative and in a non-systematic fashion. These examples of social software which provide benefits to enterprise processes, are contained services, from different providers. As part of a business process, they would most likely be integrated at differing times with no concept of linking one human agent identifying token to another across new and existing or legacy data sets. This restriction of information, produced and consumed by human agents within a system, limits the mining and analysis potential.

With enterprise adoption of such services being a la carte in nature, by leveraging a platform agnostic approach, an enterprise may add best of breed social applications and infer the digital identity across disparate legacy data sets. From this foundation, a fundamentally richer understanding of the human agents within the system may be ascertained. By creating an automated programatic digital identity resource, more interconnections between the human agents and their data may be found across the entire business process. From this aggregated social data, a deeper level of understanding in the areas of expertise and reputation may be found.

3.2 Reputation-based Message Routing and a Unified Digital Identity Resource

This subsection will present a scenario being explored currently, which will help to put the concept of a programmatically available digital identity resource in context. By providing a programmatically accessible digital identity resource via a REST based JSON/XML api, a social software system will enable analysis of the interconnected nature of the human agents acting within the system and the data which they generate. By layering human agent generated data on top of existing data, new forms of interconnectedness can be mined. This interconnected data on human agent activity will, in turn, enable more light weight ad hoc business processes to be executed in a more effective manner. Rather than

adopting an abstraction of the human agent within a system, via worklists or pattern identification [34], the system, called Reputation-based Message Routing, takes an alternate approach by providing human agent recommendations based on such analysis gained from data artefacts such as mailing lists, commit logs and bug ticketing data [19] down to the granularity of a specific human agent.

A simplified example of such analysis can be seen in the diagram below (Figure 1). An email thread and a bug report have two common human agents (agent one and two). From this initial inferred connection, the system can infer that a second email thread, which also has agent two as a participant, might be of interest to agent one. For a real use case, there would need to be many more inferred connections of commonality before such a recommendation would be made, but in the small, such an aggregation of content and agents across data sets can show interesting results. This is only possible once a unified digital identity resource is in place.

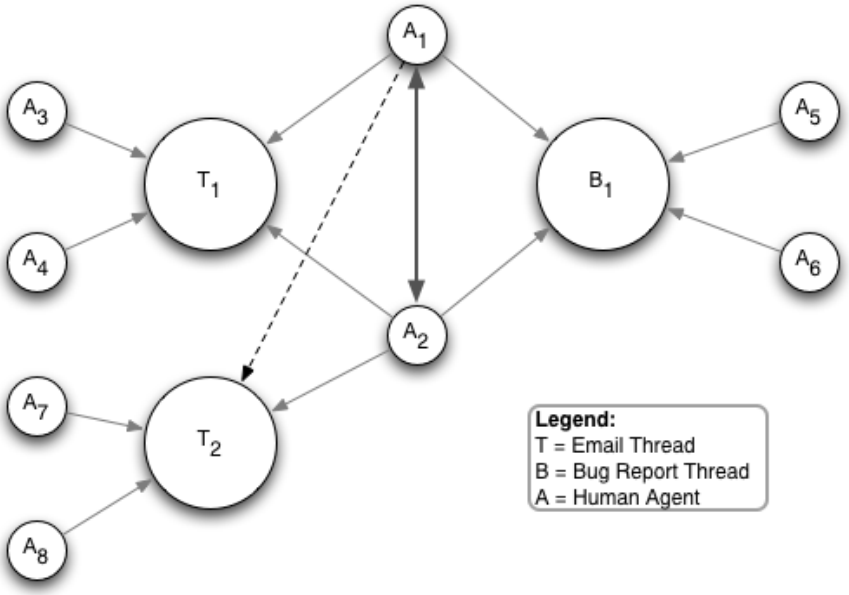

Fig. 1. Thread, Bug and Human Agents Connectivity

As well as a fully automated approach to the identification of a specific human agent within the enterprise, this approach will also be extended by enabling a lightweight mechanism for enabling human agents to correct and extend the automated identity recognition. By letting people tag people [12], as well as the machine oriented approach, a significantly richer source of identity is possible

as well as using human intelligence and expertise to correct and extend the automated identity recognition. By leveraging existing assets, this approach will solve the adoption problem and bring a greater value to the adoption of social software within the enterprise as well as providing contextual meaning to the relationship between people and between data and its creators.

4 Digital Identity Pattern Extraction (DIPE)

In the previous section, the likelihood for a wide range of disparate social software was discussed. From this, the need for an automated digital identity mechanism could bring value to a wide range of enterprise needs. As part of the Reputation-based Message Routing system mentioned above, an automated mechanism was needed in order to process a real world data set. In this section, a comparison between a previous approach and the DIPE prototype will be discussed, then a brief look at some methodological concerns. In the last part of this section, some of the concepts behind the DIPE prototype will be addressed and some early results presented.

4.1 Automation and Methodologies

In work from 2006 [3], an idea was presented for dealing with the analysis of part of the Apache project and the need for resolving identity within a mailing list. This approach was based on having a well formed RFC 2822 token, foo.bar@apache.org (Foo Bar) for example. Their approach enabled clustering of potential results via distance vectors [29] which were then hand sorted. Whilst this approach looks promising for well formed data sets, when working with a heterogeneous and legacy environment, there is the strong potential for mixed and non-well formed data as identifying tokens. There is also the strong potential for intentionally corrupted data to avoid spam harvesting.

In order to enable lightweight ad hoc horizontal macro-operation [31] between human agents, there needs to be a sound foundation for identifying specific agents in relation to context specific criteria. In the data set currently being examined, there are in excess of 3000 distinct human agents operating across three substantive data sets, the email list having over 22,000 items. The bug ticketing system has a wide variety of identifying tokens being used, from well formed RFC 2822 (foo.bar@apache.org (Foo Bar)) to just a name (Foo Bar), to an intentional corruption or obfuscated token (foo dot bar [at] apache dot or_g). By using the previous approach, a distance vector would not lead to sound results in this domain. A novel approach is needed to solve this problem.

Before discussing this approach, a short discussion of methodological concerns should be raised. When dealing with such data, there is no technique for creating an oracle from which to validate results. As the results are subjective, particularly in relation to the intentionally corrupted data, the verification of the algorithms needs to be tested by hand. Since the data sets are real world and

substantive, random sampling is the most appropriate method [37]. The other methodological factor of the DIPE algorithms is that they target high value i.e. human agents with a larger volume of interaction with the system. With this in mind, the random sampling is intentionally biased towards the top end of the Pareto distribution of the network [30].

4.2 Current Prototype for DIPE

The DIPE prototype is an entirely automated method for extracting digital identity from heterogeneous data sets. As was discussed in the previous subsection, a vector based solution, whilst appropriate in some instances, would lead to suboptimal results with this quality of data. The approach is based on multiple best effort heuristic algorithms, based on commonly used patterns within name based tokens. The approach falls into four basic stages:

- *Harvester:* In order to maintain portability, the DIPE prototype deals with openly available data sources. To make the system entirely automated, the Harvester needs to be able to retrieve data in an unsupervised manner. The Harvester intelligently backs off in order to not put undue strain on remote servers.
- *Parser with Wrappers:* The Parser is written in an extensible manner in order to add additional data formats in the future. These wrappers then add new token types to the database.
- *Classifier:* As DIPE needs to react to differing data types, the Classifier needs to detect type of tokens. Using regular expressions, it looks for common identifiers in order to apply the most appropriate algorithm. The RFC 2822 algorithm, rather than a vector approach, looks for patterns asserted by the human agent. In a conservative approach using the most likely target of high quality data (the first/last name assertion) down to commonly used separators (periods, hyphens, underscores), it attempts a first/last name extraction. The deobfusticator works in an entirely different manner. As the human agent is intentionally trying to mislead pattern recognition, the heuristic must look for common conventions. For example, replacing @ with at and . with dot. There are many more such conventions such as partitioning data with matched or unmatched tokens, brackets or square braces for example. It also employs a recursive algorithm to look for implied delineation via in place case alteration (fooBarApache for example). The deobfusticator looks for all such patterns and attempts a first/last name extraction.
- *Matcher:* The Matcher takes all data from all available data sets and simultaneously looks for matches both internally within each set, then across all sets. The internal matching tests for partial string matching in the first/last name pairs, for domain similarities and compounded words. The cross set correlation takes advantage of implied hierarchy of sets, i.e. some sets having greater importance than others. It uses a name frequency inverse population approach.

4.3 Preliminary Results

These early results are based on a real world data set and the entire set is being processed in an automated manner. There was no cherry picking of easy to deal with data types, DIPE takes real data and via the best effort heuristic algorithms attempts to cluster identities across heterogeneous sets:

Table 1. DIPE Preliminary Results

Case	Set#1	Set#2	Set#3
Distinct IDs	2748	486	32
Resolved Aliases	60%	55%	n/a
High Value	68%	67%	94%
Cross Correlated	62%	61%	91%

From these early results, it can be seen that a heuristic approach can lead to worthwhile results from legacy and real world data sets in an automated manner. These results are currently being analysed to focus on sections of DIPE to improve the confidence level of the results set.

5 Conclusions and Future Work

In this work, it has been observed that the enterprise is likely to adopt social software in a conservative, non-systematic manner. In order for enterprise to leverage the social graph to integrate human agents in business processes in a more meaningful manner, there needs to be a mechanism to create a unified digital identity resource in an automated manner. From this resource, foundations for trust and reputation can be built and this will enable business processes to have a richer source of information from which to make human agent selections. The DIPE prototype demonstrates that real world data, even intentionally corrupted data, can be mined with a relative degree of confidence in order to unify social enterprise assets. This confidence is based on the subjective interpretation of the data set (see section 4.1).

This digital identity resource forms part of a larger context, the Reputation-based Message Routing system. Building on solid automated digital identity foundation, via DIPE, RMR will perform data mining techniques on that data to derive relationships of human agents to each other and their data. By establishing many to many relationships, this basis of interconnected data provides a reputation based engine for making human agent recommendations both in the selection process and the execution of their work.

Leveraging existing enterprise data sets, by creating a pragmatic unified digital identity strategy, there is the potential for creating a speedy but not rash [8] integration of social software in Enterprise 2.0.

References

1. G Akerlof, *Market for lemons: Quality uncertainty and the market mechanism*, ideas.repec.org.
2. K. Binmore and P. Dasgupta, *Game theory: A survey*, Economic Organizations as Games (Oxford: Basil Blackwell) (1986).
3. C Bird, A Gourley, P Devanbu, M Gertz, and A Swaminathan, *Mining email social networks*, Proceedings of the 2006 international workshop on Mining software repositories (2006), 137–143.
4. J.L. Bradach and R.G. Eccles, *Markets versus hierarchies: from ideal types to plural forms*, Annual Review of Sociology **15** (1989), 97–118.
5. F.P. Brooks, *The mythical man-month*, Mass.: Addison-Wesley Pub., 1979.
6. R Burt, *A note on social capital and network content*, Social Networks (1997), 355–373.
7. K. Clugage, D. Shaffer, and B. Nainani, *Workflow services in oracle bpel pm 10.1.3*, `http://tinyurl.com/2sal77`, 3 2006.
8. TH Davenport, *Putting the enterprise into the enterprise system*, Harvard Business Review **76** (1998), no. 4, 121–131.
9. C. Dellarocas, *Analyzing the Economic Efficiency of eBay-like Online Reputation Reporting Mechanisms*, MIT Sloan School of Management (2001), 4181–01.
10. L Efimova and J Grudin, *Crossing boundaries: A case study of employee blogging*, Proceedings of the Fortieth Hawaii International Conference ... (2006), 11–24.
11. N Emler, *A social psychology of reputation*, European Review of Social Psychology (1990).
12. S Farrell and T Lau, *Fringe contacts: People-tagging for the enterprise*, Collab. Web Tagging Workshop in conj. with WWW2006 (2008), 10–100.
13. C Fombrun and M Shanley, *Whatś in a name? reputation building and corporate strategy*, Academy of Management Journal (1990).
14. D Gambetta, *Can we trust trust*, Trust: Making and Breaking Cooperative Relations, electronic edition, Department of Sociology, University of Oxford (2000), 213–237.
15. N Good, J Schafer, J Konstan, and A Borchers, *Combining collaborative filtering with personal agents for better recommendations*, Proceedings of AAAI (1999), 9166–9172.
16. T Grandison and M Sloman, *A survey of trust in internet applications*, IEEE Communications Surveys and Tutorials **3** (2000), no. 4, 2–16.
17. R Guha, R Kumar, P Raghavan, and A Tomkins, *Propagation of trust and distrust*, Proceedings of the 13th international conference on World Wide Web (2004), 403–412.
18. SR Hiltz and M Turoff, *Structuring computer-mediated communication systems to avoid information overload*, Communications of the ACM **28** (1985), no. 7, 3310–3312.
19. B. Jennings and A. Finkelstein, *Flexible Workflows: Reputation-based Message Routing*, Proceedings of BPMDS **8**, 97.
20. Y Kakizawa, *In-house use of web 2.0: Enterprise 2.0*, Nec Technical Journal (2007), 8623–8635.
21. C Li, *Blogging: Bubble or big deal?*, Forrester Research Document (2004), 9290–9392.
22. C. Locke, D. Weinberger, and D. Searls, *The Cluetrain Manifesto: The End of Business As Usual*, Perseus Publishing, 2001.

23. L Luna-Reyes, A Cresswell, and G Richardson, *Knowledge and the development of interpersonal trust: a dynamic model*, System Sciences (2004).
24. P Massa and I ITC-IRST, *A survey of trust use and modeling in real online systems*.
25. A Mehrabian, *A questionnaire measure of individual differences in stimulus screening and associated differences in arousability*, Journal of Nonverbal Behavior **1** (1977), no. 2, 89–103.
26. D Millen, J Feinberg, and B Kerr, *Dogear: Social bookmarking in the enterprise*, Proceedings of the SIGCHI conference on Human Factors in ... (2006), 8561–8580.
27. G Möllering, *The nature of trust: From georg simmel to a theory of expectation, interpretation and suspension*, Sociology **35** (2001), no. 02, 403–420.
28. L Mui, M Mohtashemi, and A Halberstadt, *A computational model of trust and reputation for e-businesses*, Proceedings of the 35th Annual Hawaii International Conference on System Sciences (HICSS'02)-Volume 7-Volume 7 (2002), 3184–3200.
29. G Navarro, *A guided tour to approximate string matching*, ACM Computing Surveys (CSUR) (2001), 31–88.
30. MEJ Newman, *The structure and function of complex networks*, Structure **45** (2004), no. 2, 167–256.
31. S Nurcan, *Analysis and design of co-operative work processes: a framework*, Information and Software Technology (1998), 143–156.
32. A Quan-Haase, J Cothrel, and B Wellman, *Instant messaging for collaboration: A case study of a high-tech firm*, Journal of Computer-Mediated Communication **10** (2005), no. 4, 00–10.
33. W Raub and J Weesie, *Reputation and efficiency in social interactions: An example of network effects*, American Journal of Sociology (1990).
34. N Russell, A.H.M. ter Hofstede, D Edmond, and W.M.P. van der Aalst, *Workflow resource patterns*, BETA Working Paper Series (2004).
35. C Shapiro, *Consumer information, product quality, and seller reputation*, Bell Journal of Economics (1982).
36. B Stvilia, M Twidale, and L Smith, *Information quality work organization in wikipedia*, (2008).
37. W Tichy, *Should computer scientists experiment more?*, doi.ieeecs.org (1998).
38. C Tinsley, K O'Connor, and B Sullivan, *Tough guys finish last: the perils of a distributive reputation*, Organizational Behavior and Human Decision Processes (2002).