A FOUNDATION FOR OPEN INFORMATION ENVIRONMENTS

Research in Progress

Parsons, Jeffrey, Memorial University, St. John’s, Canada, jeffreyp@mun.ca
Wand, Yair, University of British Columbia, Vancouver, Canada, yair.wand@sauder.ubc.ca

Abstract

Traditionally, information systems were developed within organizations for use by known audiences for known purposes. Advances in information technology have changed this landscape dramatically. The reach of information systems frequently extends beyond organizational boundaries for use by unknown audiences and for purposes not originally anticipated. Individuals and informal communities can generate and use information in ways previously restricted to formal organizations. We term applications with these characteristics open information environments (OIEs). OIEs are marked by diversity of information available, flexibility in accommodating new sources, users and uses, and information management with minimal controls on structure, content, and access. This creates opportunities to generate new information and use it in unexpected ways. However, OIEs also come with challenges in managing the semantic diversity, flexibility of use, and information quality issues arising from the range of users and lack of controls. In this paper, we propose a set of principles for managing OIEs effectively. We outline a research program to examine the potential of OIEs, the challenges they present, and how to design OIEs to realize the benefits while mitigating the challenges. We highlight our ongoing research in this area, and conclude with a call for more research on this important phenomenon.

Keywords: Open Information, Semantic diversity, Information quality, Modelling Principles.

1 Introduction

Traditionally, information was managed within well-defined organizational structures in environments that enabled and emphasized control over the contents and structure of the data (Mason and Mitroff, 1973). In such environments, data were generated, collected, and used almost exclusively within organizational boundaries with specific and well-defined uses in mind, and were generated by well-known sources that enabled a high level of oversight of the data created. We refer to these settings as closed information environments (CIEs).

Against this backdrop, rapid advances in Internet technology and dramatic increases in digital storage and processing power have radically transformed how people and organizations create, obtain and use information. These changes have enabled new opportunities for formal organizations, and have given rise to a proliferation of informal information generating and sharing communities. In such communities, data are not necessarily created or used within, by, or for, a formal organization. A well-known example is data crowdsourcing (user-generated content), in which the general public is engaged in contributing data (Doan et al., 2011).
This information transformation is marked by three main characteristics that conflict with CIEs. First, information diversity means that data can be obtained and combined from multiple independent and diverse sources. Second, information flexibility is the ability to accommodate emerging sources, users, and uses of information. Third, use (user) independence means information can be generated, accessed, and used with minimal constraints on, or control over, structure and contents. We refer to environments with these characteristics as Open Information Environments (OIEs).

Throughout the paper, we use a specific example of an OIE – namely, citizen science – to illustrate concepts. In particular, we consider a data-oriented citizen science project in which contributors provide sightings of natural history phenomena (Lukyanenko et al., 2011). CIEs avoid data diversity issues and provide controls to assure data quality. However, CIEs cannot (nor are they intended to) enable diversity of content, flexibility of generation and use, and the independence afforded by OIEs. This leads to the question: how can the potential advantages of OIEs be realized?

In this paper we propose a set of information management principles intended to support and realize the potential of OIEs. In the following, we first expand on the OIE concept and its associated opportunities and challenges. We then propose a set of requirements for OIE success, and articulate principles for achieving these requirements. We outline a research program that is studying several key issues in the design, adoption, and evaluation of OIEs, and highlight some key findings of research in progress. We conclude with a summary of the potential significance of this important emerging area.

2 Open Information Environments and Their Challenges

We define OIEs as environments in which: users have access to sources over which they may have no control; new sources of data may emerge; applications of data might change radically over time; and new uses of data might emerge. The key question is how to overcome the challenges such environments might create. To understand these challenges, we consider three (possibly overlapping) perspectives, each reflecting a different type of stakeholder. These are information contributors, information consumers, and OIE sponsors. Contributors generate data; in other words, they create the content available in information sources. Contributors can be individuals, organizations, communities, or automated applications. In a citizen science application, contributors are ordinary citizens, about whom little may be known. This group is very diverse, not under any central control, and may not know how the data will be used. Consumers are the users of OIE data. Again, they can be individuals, formal or informal organizations, or automated applications. Their challenge is to use data from disparate sources, without knowing what meaning contributors assigned to the data or even what useful information or insight can be gleaned from the available data. In citizen science, users are typically scientists. Sponsors are those who allocate resources (e.g., financial and technical) to create the underlying OIE mechanisms, and need to evaluate its success in serving the needs of some stakeholders. Sponsors might be consumers, or organizations with some relationship to consumers. Sponsors typically have concerns regarding: data quality; data provenance (related to authenticity); the ability of systems to support flexible and changing requirements; and trustworthiness of information.

---

1 The concept of OIE should not be confused with that of open data (or linked open data) (Auer et al., 2007). The latter deals with applying Semantic Web technologies around existing independent data sources to allow machines to understand, reason about, and combine independent data. OIEs are broader, as they deal with issues such as providing flexibility of data capture and use, enabling semantic diversity without central controls, and improving data quality at point of capture. OIEs may be supported by Semantic Web and linked data concepts, but are independent of these technologies.
sources. Note that we do not address here the “traditional” issues of security, safety, integrity, and privacy of data.

OIEs create unparalleled opportunities to generate, integrate, and use information, within and across formal organizations and user communities, by accommodating diverse information contributors and providing the flexibility to support their emerging needs. However, OIEs come with challenges. In particular, two major barriers can inhibit OIEs from delivering on their promise. First, flexibility of applications and diversity of users and sources creates a semantic diversity challenge, wherein different users and different sources have different views (interpretations) of the information to be shared. Users might not know the meaning or structure of data generated elsewhere. Even if they do, the differences in how the information is viewed by users and by the sources might create a major obstacle for users in attempting to use the information. Second, because OIEs do not incorporate central controls and sources might be unaware how the information will be used, the accuracy, completeness, and suitability of the information for a given use cannot be assured, creating an information quality challenge. Thus, a tension exists in OIEs between the potential benefits of flexibility and the challenges of achieving effective use and information quality. For example, in a natural history citizen science project, biologists (consumers) are typically interested in data at the species level, whereas citizens (contributors) often are unable to provide accurate data at that level of granularity (Parsons et al., 2011).

A key to understanding the semantic diversity challenge is the notion of user view (Navathe et al., 1986). A view is a conceptualization of the domain about which information is collected and used. The types of concepts in a view determine the semantics associated with the data. Users with different views might assign different meaning to the data they generate or that are available to them. Moreover, users might change their views over time. To enable a variety of information users (contributors and consumers) to engage with the same data for different purposes, an OIE implementation requires mechanisms that allow users with different and evolving views of the phenomena in a domain to access the data through these views. This cannot be addressed by traditional information management methods that base data organization on some assumed views that reflect a fixed (or slowly evolving) organization of the data (Navathe et al., 1986).

In the information systems (IS) context, a view is formalized as a conceptual model—a set of concepts (and their relationships) used to describe a domain (Batini et al., 1992). A main use of such models is for analyzing IS requirements (Wand & Weber, 2002), and therefore they are expected to reflect the views of the expected users of an IS. Semantic diversity necessitates that users be able to apply their own conceptual models to the information, independent of the model used to create the data.

To understand the motivation for the principles we will suggest for OIEs, we need to understand why traditional information environments cannot support OIEs. We argue that for OIEs to be successful, they need to satisfy three requirements:

1. **Accommodate Semantic Diversity**: As different OIE users (contributors and consumers) have different and varying conceptualizations of information, it is necessary to be able to accommodate information provided by contributors with unknown views. In addition, an OIE should accommodate the needs of a variety of information consumers whose views differ from each other and from those of information contributors.

2. **Ensure Information Quality**: With varying conceptualizations of information, and lack of central control, it is challenging to ensure information quality (IQ) in OIEs using accepted dimensions

---

2 Information quality is a multidimensional construct, with key dimensions including accuracy, completeness and timeliness (Pipino et al., 2002; Wang and Strong, 1996).
such as accuracy and completeness. This is particularly important when contributor conceptualizations do not match those of information consumers.

3. **Enable Flexibility**: The notion of openness in an OIE implies a need to accommodate rapid and unanticipated changes in input sources, access needs of consumers, and information use.

These requirements are not unique to OIEs, and have been recognized and dealt with in traditional applications, such as those used in formal organizations. However, traditional approaches sought to address these issues based on four underlying assumptions:

1. **Premise of Central Control**: Central control is possible and desirable. The type of contents and the structure of the information can be pre-determined and kept relatively fixed. Users can be told what information is available and how it is to be accessed. Controls on data input can be implemented and enforced. Information can be organized and presented according to known uses.

2. **Premise of Known Sources**: Information sources are well known and the kind of information available is known and well-defined.

3. **Premise of Known Uses**: Information consumers and consumption patterns are known.

4. **Premise of Stability**: The pace of change is slow enough that each change can be accommodated in time. For this purpose, “in time” means “before the next change is required”.

Within CIEs, these principles are enforceable and promote efficiency and quality with respect to predetermined uses. However, they cannot accommodate the general needs of diversity, flexibility, and independence of OIEs.

Next, we propose principles that can serve as a foundation for OIEs.

## 3 Proposed OIE Principles

We now describe the principles we propose for supporting OIEs. Our purpose is to provide for the three main aspects of diversity, flexibility, and use independence.

We begin by defining some concepts. First, we use the word *user* in the most general sense. A user can be an individual, a formal or an informal group or organization, or an automated application. A user may be engaged with an OIE as a contributor (e.g., a citizen scientist), a consumer (e.g., a biologist), or in both roles. Second, we use *data* to refer to representations of characteristics of *phenomena* in domains represented by the data (these domains can be real, imaginary, or combinations) and *information* to refer to data along with the context or interpretation provided by a contributor or consumer. For example, a citizen scientist may report an observed bird with yellow wings (data). A biologist using this data may be interested in the extent to which the observed attribute is useful in narrowing the range of possible species. Finally, we use the term *engage with data* to indicate both generation and use of information. We term a statement about a characteristic of a phenomenon a *property*. Properties are manifested in data as *attributes*. We term a single statement or a set of closely related statements (e.g., observed concurrently about the same phenomenon) an *instance*. In a citizen science context, an instance can be viewed as a set of statements about an observed organism reflecting attributes observed by the contributor.

Consider the flexibility and diversity required in an OIE. First, diversity implies that, at a given time, different users may apply different conceptual views to the data. Therefore, the same data need to be accessible with different views. Second, flexibility entails that, over time, new uses and new users may emerge, new data may be contributed, and existing data may be used in new ways. This means that the conceptual views applied to the data cannot be predicted. Specifically, when data are created, the creator might not know all possible future uses of the data even if current uses can be assumed to be known. Thus, the need to support different and evolving views leads to the following principle:
**Principle 1 (View on Demand):** OIEs should support the creation of information views at the point of use.

Since views are defined in terms of logical structures, we also term this principle *structure on demand*. Consider now that a conceptual model is usually defined in terms of *classes* – reflecting domain concepts – and their possible relationships. Principle 1 implies that data about instances of the phenomena in the application domain should be assigned to classes only when the information about them needs to be used. Implementing this principle would enable different information users, at different times, to classify data as needed.

Consider now the case of traditional information management. Data are usually modeled to reflect some expected uses. Indeed, a fundamental tenet of requirements engineering is to identify the users and their information needs (Potts et al., 1994). Even agile development methods assume users are available when a system is developed (Cockburn and Highsmith, 2001). Moreover, and critical to our discussion, present technology usually links the logical view of the data (the information model) to the way the data are actually organized. This is manifested by various approaches to translating logical data models (e.g. Entity-Relationship diagrams) to specific representations in terms of data structures (notably relations in a relational database). In short, in the traditional ways of organizing information, classes are viewed as necessary and stable “containers” from which all information is retrieved (Kent, 1978). This approach is intended to contribute to efficient processing and helps retain a (predetermined and fixed) context of contributed data. However, from an OIE point of view it is limiting. First, use independence is inconsistent with imposing a specific view on users. Second, the traditional approach does not readily accommodate multiple views (see research on how to integrate different users’ views, e.g., Spaccapietra and Parent, 1994) or rapidly evolving views (Lerner and Habermann, 1990). Third, given that in an OIE data might be created without knowing its possible uses, this approach cannot even be feasible for OIEs in general. We therefore propose the next principle of OIEs:

**Principle 2 (Instance Independence):** Information about individual phenomena (instances) should be stored independent of any classification (Parsons and Wand, 2000).

There is a stark contrast between this principle and the common paradigm of information design, which begins by defining classes of information (Song et al., 2004). For example, Tsichritzis and Lochovsky (1982) define “data item” in a data model as a member of an a priori category, so that “data that do not fall into a category have either to be subverted to fall into one, or they cannot be handled in the data model” (p. 8).

To provide an analogy to this principle, consider the move to the relational database model from the hierarchical and network models. This transition can be construed as removing from the data structures links that reflected specific applications or intended uses of the information. These links contributed to efficiency, but confounded the data structures with specific, fixed uses. Using the data required knowledge about the particular structures. Moreover, uses that did not follow the embedded structures required major effort in either reorganization, or in processing. In contrast, links in the relational models could be created “on demand” (using joins) to reflect the application needs. Conceptually, our approach suggests a similar step – remove from the organization of the data fixed “containers” for data instances, because such containers reflect specific anticipated uses of data.

Unbundling storage of data (instances) from classification structures that reflect specific views can support multiple and evolving views without the need to reorganize the data or to apply complex processing operations. However, one issue remains unresolved. A conceptual view is defined in terms of classes. There are different approaches to classification (Smith and Medin, 1981). In particular, in the “classic” approach a class is defined in terms of a set of necessary and sufficient properties. Other approaches exist (notably “prototype” and “exemplar” views). However, all approaches are based on the fundamental notion of a property. This implies that, to support diversity of views of the same data,
a way is needed to reconcile representation of properties across users. To accomplish this, we propose the third OIE principle:

**Principle 3 (Data abstraction):** Properties can be arranged in an abstraction hierarchy. In such a hierarchy, lower level dissimilar properties may be reconciled as higher level similar ones (Parsons & Wand 2003).

An example for property abstraction is a business application is the notion of “preferred customer”. In independent data sources, the two distinct properties “loyalty club members” and “has made large purchases” can be abstracted and integrated in a more general notion of “preferred customer”.

Next, we outline a program of research to determine whether and how OIE principles contribute to enabling semantic diversity, flexibility, and information quality.

## 4 Research Opportunities

The proposed research is based on the four fundamental premises described above:

1) OIEs afford opportunities to generate, integrate, and use information by accommodating diverse information contributors and supporting emerging needs of diverse information consumers. However, these opportunities come with the potential challenges of *semantic diversity* and *information quality*.

2) Traditional information management approaches cannot realize OIE benefits without incurring the costs. We propose that our set of OIE principles can overcome this trade-off.

3) There are two main types of situations where information is created and used: well-defined (formal) organizations and less structured (informal or semi-formal) heterogeneous user communities.

4) OIEs involve the three roles of information contributor, information consumer, and project sponsor.

The research program examines how OIEs can be made effective by applying the principles of our theoretical approach. Our studies are exploring two of the three types of roles (contributors and consumers) in an informal setting (user communities). *We hypothesize that, by eliminating the constraints associated with CIEs, OIEs based on our principles will lead to greater information quality, improved ability to use information in unanticipated ways, and greater user satisfaction.* We will examine these questions using field experiments to study effects on information quality and user satisfaction, and a case study in the citizen science domain to study the contribution of an OIE to using information in ways not anticipated when the project was initiated.

The primary *independent variables* are based on the OIE principles. In particular, we examine the impact of *Information Flexibility*, defined as the degree of freedom a contributor or a consumer has to provide or access data in ways that best reflect his/her view of the domain about which data are collected or used. Information Flexibility from a contributor’s perspective will be manipulated in terms of *Input Flexibility* (IF) – defined as a four-level construct: (1) **no flexibility** – the information structure and interface reflect the requirements of a specific application in terms of a given collection of classes (each specifying a fixed set of information attributes); (2) **low flexibility** – the information structure and interface reflect a set of generic or “basic level” (Rosch, 1978) classes to which the contributor can add information attributes; (3) **moderate flexibility** – the information structure and interface reflect a specified list of information attributes, independent of classes, from which the contributor can select; and (4) **full flexibility** – the information structure and interface allow the contributor to add new attributes as well as select from an available list. For levels 3 and 4, users might define their own classes. This variable is operationalized and manipulated by designing a separate user interface to enable each condition. The interface structure reflects an underlying class-based database structure (levels 1 and 2) or data attributes (levels 3 and 4).
Information Flexibility from a consumer’s perspective is manipulated via Information View Control (IVC) – defined as a four-level construct: (1) No user control – information structures (classes) are given; (2) Low user control – information is organized in basic-level classes, which the consumer can modify by adding attributes (from a given list) or by combining classes; (3) Moderate user control – information is structured only in terms of attributes, and consumers can form a set of information classes by combining attributes; and (4) Full user control – information is structured in terms of attributes, and consumers can specify new attributes that are generalizations of attributes with similar meaning. Users can also form classes that reflect their view, including the newly-defined attributes.

This variable will be operationalized by designing interfaces that enable consumers to explore the information structures or add to them according to each of the above levels.

The dependent variables include objective measures of information completeness, information accuracy, ability to obtain the required information, and ability to discover new useful classes, as well as subjective measures of user satisfaction with the interface, perceived ease of use and perceived usefulness. The combination of objective and subjective measures will help provide a comprehensive picture of the benefits of OIEs. We will use validated measures of constructs where available (e.g., perceived ease of use and perceived usefulness (Davis, 1989)). Otherwise, we will develop appropriate measures, following an approach used previously to compare conceptual models (Parsons & Wand, 2008).

In addition, we will use interviews and observations to explore problems that are attributable to the limitation of traditional approaches.

We have identified an informal organization (user community) in which to conduct this work – a crowdsourcing platform supporting citizen science. We have access to a citizen science project focused on the flora and fauna of a geographic region in Canada. The project provides a rich environment with a diverse range of contributors (from practicing ecologists to people with little ecological knowledge), a clear “standard” on the consumer side (biological taxonomy) that does not match the conceptual model of most contributors, and a high potential for unanticipated uses of data.

Next, we briefly describe some preliminary research we have done in the area of informal organizations.

5  Experience with the Principles in an Informal Organization

Citizen science is the engagement of members of the general public to support scientific research (Silvertown, 2009). In particular, citizen science can be used to enlist the public in collecting research data. Data collection in citizen science exemplifies many of the characteristics of OIEs in informal organizations: (1) citizen scientists (information contributors) may have very different levels of knowledge of the scientific domain (from novice to expert), they are volunteers and therefore the project owner has limited ability to control or influence participation, and they may have limited insight into how the data will be used; (2) scientists who are the information consumers have difficulty specifying in advance what useful information or insight can be gleaned from the available data (Dickinson et al., 2010; Wiersma, 2010); and (3) scientists who are project sponsors have concerns about the quality of data provided by amateurs (Cohn, 2008; Dickinson et al., 2010; Gura 2013).

Several studies have been conducted, focusing thus far on information quality from both the contributor and consumer perspectives. In particular, we have examined the effect of imposing predetermined and fixed classification on input data on the accuracy and completeness of data. We have found that requiring that data be entered in terms of scientifically useful categories (species of plants and animals) had a severe negative effect on the accuracy of data. At the same time, although citizen scientists were generally unable to provide accurate data at scientifically predetermined levels, when given the opportunity and flexibility, they were able to report considerable additional data (in
terms of attributes) about the phenomena they observed. Currently, we are investigating the extent to which this data can be useful to scientists, even though it was not collected in terms of the classes the scientists had a priori deemed useful. As well, we have experimented with a prototype implementation of an OIE based on Semantic Web technologies. The data instances were stored in an RDF data base (Klyne and Carroll, 2013), and could be accessed based on a flexible hierarchy of attributes and a dynamic set of classes defined using the Web Ontology Language (OWL) (http://www.w3.org).

We have also experimented with storing data in terms of instances and attributes and examined issues in the design of user interfaces when contributors do not provide data in terms of fixed classes.

Results of these studies will be ready for presentation at the conference.

6 Conclusions

Emerging technological, organizational, and social environments provide exciting opportunities to use information in new ways that can radically increase the value of information to users, organizations, and society. These opportunities come with two main challenges – semantic diversity and information quality. Our OIE conceptualization and principles provide a way to benefit from OIEs while overcoming the challenges. Specifically, the proposed research will lead to: (1) better understanding of the limitations of closed applications and the potential of OIEs; (2) better understanding of how to implement information flexibility, and evidence for its usability and usefulness; and (3) better understanding of how information quality challenges can be overcome in OIEs. We believe OIEs can provide a modern information infrastructure for supporting open information contribution and access.

References


