Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Publishing metadata of geospatial indicators as Linked Open Data: a policy-oriented approach

Diederik Tirry
KU Leuven/SADL
Celestijnenlaan 200E
Leuven, Belgium
diederik.tirry@sadl.kuleuven.be

Ann Crabbé
KU Leuven/SADL
Celestijnenlaan 200E
Leuven, Belgium
ann.crabbe@sadl.kuleuven.be

Thérèse Steenberghen
KU Leuven/SADL
Celestijnenlaan 200E
Leuven, Belgium
therese.steenberghen@sadl.kuleuven.be

Abstract

Geospatial indicators are becoming increasingly important for governments in monitoring and underpinning policy planning and political decision making. Currently, the discovery, viewing and sharing of these indicators is often made possible through geoportals that are developed according the concepts of Spatial Data Infrastructures (SDIs). However, this type of 'business information' exceeds the scope of traditional SDIs that solely focus on the common spatial aspects constituting a generic location context. The concept of an 'augmented' SDI adopting Linked Data principles reveals meanwhile much potential in integrating disparate reference and non-spatial business data but requires a formal revision of underlying standards. In this study we propose an alternative and policy-oriented viewpoint for publishing geospatial indicators as Linked Open Data. Focussing on metadata, we have elaborated a profile of the Data Catalog Vocabulary (DCAT) for describing geospatial indicators, including additional information on the related policy assessments, spatial characteristics, the provenance, and the measurement variables and dimensions of indicators. By implementing the vocabulary in an existing monitoring system it allows us to discuss the benefits and drawbacks of this approach.
Keywords: Linked Open Data, Spatial ontologies, data catalog vocabulary, policy support

1 Introduction

Efficient and effective governance requires reliable knowledge about the current situation, the underlying driving forces, and the consequences and effects of strategic policy plans. For policy makers, the development of integrative monitoring systems is vital in order to process the multitude of information and measure the execution and outcomes of a policy program across time [3]. It is also generally recognized that the use of geospatial indicators in particular can lead to important insights in support of policy and decision making [19].

The 'Spatial Monitor Flanders' and 'Traffic Safety Monitor Flanders' are two examples of monitoring systems that facilitate a multi-level, integrative framework for collecting, publishing and maintaining the most relevant spatial indicators in these policy domains [4,17]. For both monitoring systems the concept of an SDI [8,12] was introduced earlier to connect the scattered and isolated geospatial indicators and create interoperable web services for the discovery, viewing and exchange of relevant information.

Whilst an SDI is intended to enable the access, retrieval and dissemination of geospatial information, the scope of an SDI encompasses solely common spatial aspects constituting a generic location context and therefore does not target specific applications, such as the publication of domain-specific spatial indicators via custom monitoring platforms [18]. When deploying both monitoring systems conform the SDI principles and components, a discrepancy arose between the supply of geospatial indicators and the expectations of policy makers, often less technical in nature. Therefore, the limited scope of SDIs was gradually considered as a major barrier to unlock the full value of geospatial indicators within the policy cycle.

The aim of this research is to bridge the gap between the geospatial community and policy makers by exploring how Linked Open Data (LOD) can be applied in the context of exchanging geospatial and policy-relevant indicators. In this paper we focus in particular on the metadata of geospatial indicators and present a policy-oriented approach for publishing them in the semantic web. The approach relies on the development of a new profile of the W3C Data Catalog Vocabulary (DCAT) to integrate additional metadata elements that are specific and adequate to geospatial and policy-relevant indicators.

The remainder of this paper is structured as follows: first we briefly introduce Linked Data principles and provide an overview of related research. A methodology for developing and applying a vocabulary suitable for describing geospatial indicators is presented in section 3. In section 4 we clarify the benefits and drawbacks of our approach. Conclusions and future research will be discussed in the last section of this paper.

2 Linked Open Data and SDI

The term Linked Data refers to a set of good practices for publishing and connecting structured data in the semantic web, also called the 'web of data' [2]. The notion of Linked Data is underpinned by four core principles introduced by Tim Berners-Lee in his Web architecture note on Linked Data [1]: 1) use Uniform Resource Identifiers (URIs) as reference points, 2) use dereferenceable URIs so that people can look them up, 3) encode the data in the machine-readable Resource

Description Framework (RDF) so they can be queried with the RDF query language SPARQL, 4) include links to other data sources enabling the discovery of related items. As both public and private sector have started to embrace open access and open data policies, the label Linked 'Open' Data (LOD) is now increasingly used referring explicitly to the publication of Linked Data under an open license [7].

LOD provides a new opportunity to study the use and exchange of geospatial data and information in a distributed environment, as well as to re-examine the role of SDIs implementing a service oriented architecture. In addition, the underlying semantic web technologies of LOD offer several benefits to organize the data itself on the Web and thereby using the Web as a global information space.

The use of semantics was first introduced into GIS to enable integration of disparate sources in a seamless and flexible way based on their semantic value and regardless of their representation. The generation and use of ontologies was considered as a method to provide the users with explicit information about the embedded knowledge of the information system thereby enhancing the classification process of various sources of data [6].

Triggered by the success of the LOD community, research recently shifted towards exploring the use of LOD in SDIs. Schade and Cox applied the Linked Data approach to classical SDIs and concluded that SDI concepts and Linked Data principles do no exclude but rather complement each other [15]. Different solutions were proposed to augment SDIs with LOD and improve remaining issues related to cross-community communication and cooperation [16]. At the metadata level Lopez-Pellicer et al. proposed a Linked Data frontend for CSW as a solution for publishing metadata repositories on the Web [10]. Also Reid et al. explored alternative options to publish geospatial metadata as RDF, from 'crosswalking' through well-known vocabularies such as Dublin Core, to RDF generation direct from a relational database [13]. Within the GLUES SDI project, LOD principles and technologies were applied to existing web feature services (WFS) and sensor observation services (SOS) in order to produce RDF representations of service metadata and of respectively features and observations [14]. While the abovementioned studies target individual components, Janowicz et al. presented a shared and integrative Semantic Enablement Layer that comprises a Web Ontology Service for managing ontologies and a Web Reasoning Service for integrating reasoning functionality within SDIs [9].

The concept of augmenting SDIs still faces many challenges, especially towards further elaboration and implementation. First, with regard to geospatial metadata, many of the abovementioned approaches propose well-known vocabularies such as Dublin Core terms. However, these approaches will be partially or fully overtaken if the Open Geospatial Consortium (OGC) and ISO/TC211 committee define themselves a set of Linked Data Vocabularies, hereby following the recommendations of the Delft Report on Linked Data [11]. Secondly, the software infrastructure required to produce and process geospatial Linked Open Data within an augmented SDI is currently limited to stand-alone initiatives and has not reached yet full maturity. Last but not least, most approaches in augmenting SDIs are focussed on leveraging the existing infrastructure in terms of integrating semantics for reference data, unfortunately ignoring the opportunity to establish a common ground for geospatial data and derived products such as monitoring (geospatial indicators) and reporting information.

In summary, the concept of augmented SDIs reveals a lot of potential in connecting the SDI community and the semantic web. However, current implementations are limited to pilots and sharing best practices, waiting on a formal revision of current SDI standards and transformation of existing models to RDF. Consequently, keeping an SDI-based architecture for indicator-based monitoring would impede the publication of geospatial indicators as LOD in the semantic web.

The aim of this research is to explore a new approach for publishing geospatial indicators as LOD, enabling the integration with non-spatial linked data. In the next section we propose a new pragmatic solution to publish geospatial indicators in the semantic web.

3 Methods

For publishing metadata of geospatial indicators, following patterns would be considered according the augmented SDI approach. First, existing metadata can be converted to RDF using an RDF-izer and stored in an RDF triplestore or as static RDF files. Next, the metadata is published on the web using a web server or via a Linked Data interface. Another option is to apply a Linked Data wrapper to access a catalog web service (CSW) and expose a metadata catalogue as Linked Data. Though, both patterns require that all SDI standards fully adopt the Linked Data principles.

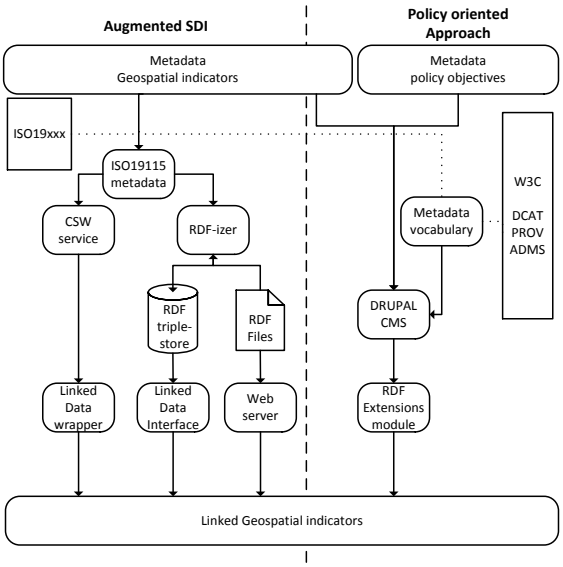
The pattern we propose is inspired by an opposite perspective on integrating metadata of geospatial indicators and Linked Data. Instead of augmenting standards from the SDI we directly select and re-use existing vocabularies that are already well-known and frequently used for describing catalogs within the Linked Data community. By extending these vocabularies with additional metadata elements, we can include information about the spatial characteristics, the policy objectives that are monitored, and the specific measures and dimensions of the geospatial indicator.

The reasoning behind is that geospatial indicators should not necessarily be described applying the ISO19115 standard because derived thematic data are considered out of scope for SDIs. Hence, we could immediately model the metadata starting from existing Linked Data specifications and seamlessly integrate our catalog of geospatial indicators with other data catalogs that are published as Linked Data. Figure 1 presents both patterns.

For the development of a vocabulary we combined a bottom-up approach, based on a use case derived from the Spatial Planning policy in Flanders, with a top-down one, analyzing the relevant semantic vocabularies. The use case helped in identifying the requirements for describing a policy-relevant geospatial indicator, whereas the review of Linked Data vocabularies provided insights into the potential eligibility of existing vocabularies.

Once the vocabulary was elaborated, it was implemented in a geospatial content management system (CMS) in order to have our target audience (i.e. policy makers) use it.

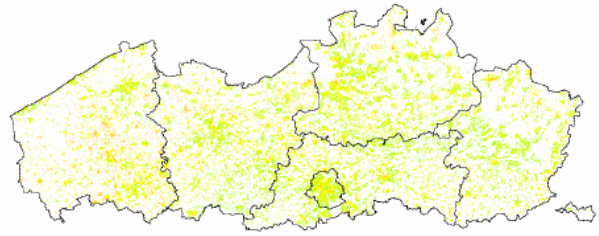
Figure 1: Augmented SDI Linked data publishing pattern (left) compared to policy-oriented approach (right)



3.1 Case study: Multi-level monitoring

Our use case involves the monitoring of the 'Green Infrastructure' for recreational purposes in Flanders. 'Green Infrastructure' is a strategically planned network of natural and semi-natural areas with other environmental features designed and managed to deliver a wide range of ecosystem services[1]. The concept is increasingly recognized by spatial planning authorities as a valuable approach for solving urban and climatic challenges. As the benefits and functions of Green Infrastructure are numerous, we focused on one single application only i.e. the role of Green Infrastructure for recreational purposes. Typical indicators used in a monitoring context are: the general provision of green space, the proximity of green space, the available green space for recreational purposes per person, demand for green space etc... Figure 2 shows an example of a typical geospatial indicator that is monitored at the regional level.

Figure 2: Proximity of green space to place of residence



Source: Natuurrapport Vlaanderen, NARA 2009 [5]

A comparison, however, between a regional and a local monitoring system revealed many semantic differences between the published indicators. We briefly describe the most important types of semantic heterogeneity among the published metadata:

No uniform metadata scheme: Each monitoring system implemented its own metadata schema to describe indicator properties, policy objectives and policy assessments. We determined significant differences in terminology and granularity of meaning.

Use of free-text fields: The ability to provide unstructured information via free-text fields for properties such as provenance, quality and relevance leads to fine-grained knowledge. However, these type of fields are prone to inaccurate information and content mismatch, because it highly depends on the author's competences and willingness to describe these properties in a correct way.

Heterogeneous classifications: Each monitoring system is using its own classification schema to categorize indicators. Whereas the regional monitoring system orders indicators according the concept of ecosystem services, the local monitoring platforms applies their own custom classification schemas. Therefore it is impossible to make a seamless integration between both platforms.

To resolve semantic heterogeneity between the two monitoring platforms we propose the introduction of three semantic components: the definition of an ontology, the adoption of controlled vocabularies and the use of taxonomies.

An ontology allows us to represent the concept of a geospatial indicator in terms of classes and properties that are applied in policy monitoring. The definition of controlled vocabularies enhance the semantic interoperability as the use of free-text is largely reduced to passively recognize a (hierarchical) list of terms as a shared context. Finally, the use of domain-specific taxonomies enables the integration of different types of indicators about the same subject e.g. Green Infrastructure.

3.2 Vocabularies

For the selection of semantic vocabularies we considered the following criteria: 1) a strong user community, 2) stable and open, 3) available in RDF, 4) adequate for our case, 5) unambiguously documented and 6) specific enough to describe indicators in sufficient detail. After a screening of existing Linked Data vocabularies, we concluded that the W3C DCAT[2] vocabulary partially suits our needs. DCAT is an RDF vocabulary designed to facilitate interoperability between data catalogs published on the Web. Hence, it supports the monitoring of indicators in different catalogs and by different government bodies. DCAT makes extensive use of terms from the Dublin Core vocabulary, which is well-known and supported by a broad community. Furthermore, it integrates the SKOS[3] vocabulary, enabling the creation of concept schemes for representing policies, structuring

[1] Green Infrastructure (GI) COM/2013/0249 final

[2] http://www.w3.org/TR/vocab-dcat/

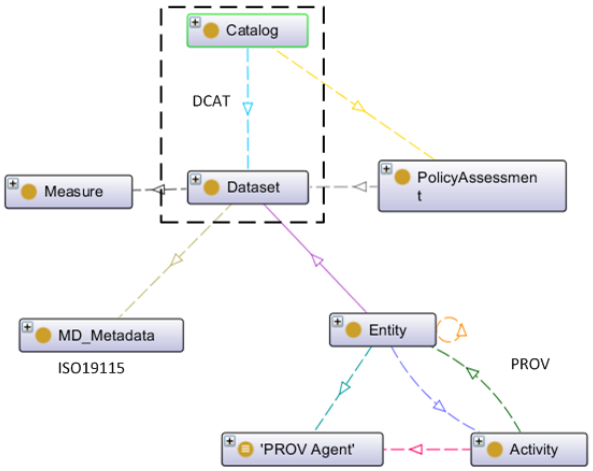
[3] http://www.w3.org/2004/02/skos/

frameworks such as the ecosystem services typology and representing an indicator typology. In the next section we discuss how DCAT can be extended to meet the remaining requirements.

3.3 DCAT-SM vocabulary

In order to meet the remaining requirements of an indicator-based monitoring system, we propose to extend the DCAT ontology and add capabilities to describe policy assessments, spatial characteristics, provenance information, and measurement information as depicted in Figure 3.

Figure 3: Extension of the DCAT vocabulary (extracted from the Protégé Ontology editor)



The result is called DCAT-SM (Data Catalog Vocabulary for Spatial Monitoring). It is developed as a profile of DCAT that includes additional information on:

- Policies: Policy assessments can be described and linked with one or more geospatial indicators. Assessments can be structured according a user-defined taxonomy (e.g. policy objectives) and linked to references and web pages that provide additional details.
- Spatial characteristics: Metadata elements describing the reference system, the resolution and the spatial representation type were extracted from the ISO19115 standard and modelled as RDF classes and properties.
- Provenance: In the context of a monitoring system it is key to understand how geospatial indicators have been calculated. The PROV ontology allows for describing provenance using structured text and/or a graphical representation of the calculation process. Via the entity class of PROV a link can be established to the core reference dataset where the indicator is derived from, avoiding the duplication of metadata elements.
- Measurements: An additional class allows to precisely describe the spatio-temporal dimensions, the thematic dimensions, the measure variables and the units of measurement. This class is indispensable for managing time series and different spatial representations of geospatial indicators. For example, the proximity of green space can be processed and represented using different reference units such as administrative regions or 1km grids.

Besides the definition of classes and properties, the DCAT-SM ontology also prescribes a series of additional classification schemes to better accommodate the Spatial Planning and Road Safety policy context, to adopt the Flemish 'Open Data' licensing framework and to include an indicator typology enabling the distinction between input, output, outcome and impact indicators.

3.4 Vocabulary implementation

The Spatial Monitor Flanders and Traffic Safety Monitor Flanders have been deployed earlier as a geospatial Content Management System (CMS) based on Drupal and integrated with Openlayers, Geoserver and PostGIS to enable geospatial capabilities such as viewing and downloading geospatial indicators.
The DCAT-SM vocabulary has been implemented by transposing each class to a Drupal content type (i.e. pre-defined collection of data types) and each property to a corresponding field type in Drupal. The content type interface allows the users to easily create and edit metadata records of indicators conform the proposed specification.
In addition Drupal has been extended with two existing Drupal modules i.e. 'RDF Extensions' and 'Restful Web Services', hereby providing extra APIs to create RDF representations of metadata records in various serialization formats such as RDF/XML, N-Triples and Turtle.

4 Discussion

Despite the potential of augmented SDIs, the SDI community is struggling with the realization of a common agreed approach for integrating SDIs and Linked Data. A significant issue is the identification of core vocabularies and a methodology how to construct mappings and transform existing metadata (and data) to RDF.
With this study we propose a different approach on the issue of sharing geospatial metadata and purposefully adopted an opposite perspective i.e. integrating Linked Data and SDI by extending Linked Data vocabularies. We try to sum up the most important benefits and drawbacks of this approach. Our approach offers the following advantages :

1. Seamless integration with 'Open Data' Catalogs: Due to the common DCAT vocabulary, catalogs listing geospatial indicators can easily be integrated in the network of emerging 'Open Data' portals.
2. Policy-oriented: The proposed DCAT-SM profile is intended for policy-makers and allows for making indicator-based assessments for any policy domain.
3. Usability: Implementing the vocabulary in an operational CMS exerts two beneficial effects on usability. First, the use of forms allows non-technical users to effortless create metadata records based on the underlying vocabulary. Secondly, the CMS offers high flexibility in

the appearance of policy assessments and geospatial indicators.

4. Accessibility: Additional APIs enable multiple representations (HTML and RDF serializations) and ensure that the content is accessible to different types of users.

Potential drawbacks of our approach are:

1. Isolation from SDIs: the suggested approach is based on the use of the DCAT vocabulary and therefore only partly relies on ontologies derived from ISO19115, disregarding most of the comprehensive schema for describing geographic data. It entails a shift away from SDIs towards the 'open data' community.
2. Narrow scope: In this study an empirical approach to publish metadata as Linked Data has been elaborated, i.e. supporting policy makers with a catalog that structures policy assessments and geospatial indicators. However, a more generic framework including formal extension patterns is indispensable to align and maintain interoperability with current Open Data portals. Ultimately, we consider such a framework as complementary to existing initiatives such as CKAN[4] in order to create catalogs that are fit-for-purpose (e.g. supporting spatial planning policy) and that are embedded in a contentful environment.

5 Conclusions and outlook

The concept of augmented SDIs reveals a lot of potential in connecting the SDI community and the semantic web but requires a formal revision of underlying standards and a transformation of existing models to RDF. In this study we propose an alternative and policy-oriented viewpoint for publishing metadata of geospatial indicators as Linked Open Data. We have established the DCAT-SM vocabulary for describing disparate geospatial indicators, including additional information on the related policy assessments, spatial characteristics, the provenance, and the measurement variables and dimensions. The specification is conceived as a profile of the DCAT vocabulary and is therefore compatible with other catalogs that have applied this RDF vocabulary.

This approach should be considered as a pragmatic and lightweight solution to bridge and integrate spatial thematic data with non-spatial Open Data repositories. With this alternative viewpoint, we also intend to contribute to the challenges on the adoption of Linked Data for geographic information.

Future research will focus on publishing the data itself as Linked Open Data, by exploring the suitability of GeoSPARQL and the RDF Data Cube vocabulary for this specific type of data i.e. geospatial indicators. Simultaneously, we also intend to widen the scope of the current approach in order to establish a more generic and formal framework for describing and distributing geospatial thematic data as Linked Open Data.

[4] http://ckan.org/

References

[1] T. Berners-Lee. Linked Data – Design issues, 2006. Available at http://www.w3.org/DesignIssues/LinkedData.html.

[2] C. Bizer, T. Heath, and T. Berners-Lee. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, vol. 5, no. 3, pp. 1–22, 2009.

[3] H. T. Chen. *Practical Program Evaluation: Assess and Improve Program Planning, Implementation, and Effectiveness*. Thousand Oaks, CA: Sage, 2005.

[4] B. Debecker, T. Steenberghen and P. Jacxsens. Spatial Monitor Flanders: Managing spatial data in support of policy making. In *Innovations in Sharing Environmental Observation and Information, Proceedings of the 25th EnviroInfo Conference*, pages 955–966. Shaker Verlag, Marburg, 2011.

[5] M. Dumortier, L. De Bruyn, M. Hens, J. Peymen, A. Schneiders, T. Van Daele & W. Van Reeth (red.). Natuurverkenning 2030. Natuurrapport Vlaanderen, NARA 2009. Mededeling van het Instituut voor Natuur- en Bosonderzoek, INBO.M.2009.7, Brussel.

[6] F. Fonseca, M. Egenhofer, P. Agouris and C. Câmara. Using Ontologies for Integrated Geographic Information Systems. Transactions in GIS 6(3), pp. 231-257, 2002.

[7] C.P. Geiger and J. von Lucke. Open Government and (Linked) (Open) (Government) (Data), *Journal of e-Democracy and Open Government*, vol. 4(2), pp. 265–278, 2012

[8] R. Groot and J. McLaughlin. *Geospatial Data Infrastructures.* Oxford, Oxford University Press, 2000.

[9] K. Janowicz, S. Schade, A. Bröring, C. Keßler, P. Maue, and C. Stasch. Semantic Enablement for Spatial Data Infrastructures. *Transactions in GIS* 14(2), Blackwell Publishing, pp. 111-129, 2010.

[10] F. J. Lopez-Pellicer, A.J. Florczyk, W. Rentería-Aguaviva, J. Nogueras-Iso and P.R. Muro-Medrano. CSW2LD: a Linked Data frontend for CSW. In *II Iberian Conference on Spatial Data Infrastructures*, Institut Cartogràfic de Catalunya, 2011

[11] F. J. López-Pellicer, L.M. Vilches-Blázquez, F.J. Zarazaga-Soria, P.R.. Muro-Medrano, and O. Corcho. The Delft Report: Linked Data and the challenges for geographic information standardization. *Jornadas Ibéricas de Infraestructuras de Datos Espaciales* (JIIDE 2011), Barcelona, 2011.

[12] D. Nebert. *Developing Spatial Data Infrastructures: The SDI Cookbook.* Version 2.0, Global Spatial Data

Infrastructure Association, Technical Working Group Report, 2004.

[13] J. Reid, W. Waites and B. Butchart. An Infrastructure for Publishing Geospatial Metadata as Open Linked Metadata. In Proceedings of AGILE 2012 International Conference on Geographic Information science, Avignon, 2012

[14] M. Roth and A. Bröring, editors. *Linked Open Data in Spatial Data Infrastructures*. Available at https://wiki.52north.org/pub/Projects/GLUES/2012-09-10_LoD_SDI_White_Paper_MR_AB.pdf

[15] S. Schade and S. Cox. Linked Data in SDI or How GML is not about Trees. In *Proceedings of the 13th AGILE International Conference on Geographic Information Science - Geospatial Thinking*, Guimarães, 2010.

[16] S. Schade, C. Granell, L. Díaz. Augmenting SDI with Linked Data. In *Proceedings of the Workshop on Linked Spatiotemporal Data*, GIScience, 2010.

[17] D. Tirry and T. Steenberghen. Towards a semantic-driven spatial monitoring framework for Road Safety. *25th ICTCT workshop*, Hasselt, 2012.

[18] K. Tóth, C. Portele, A. Illert, M. Lutz and M. N. De Lima. *A conceptual model for developing interoperability specifications in Spatial Data Infrastructures.* JRC Reference reports, Ispra, 2012.

[19] I. Williamson, A. Rajabifard and M.A.F. Feeney, editors. *Developing Spatial Data Infrastructures: From Concept to Reality.* Taylor and Francis, London, 2003.