Consistency Assessment Between Multiple Representations of Geographical Databases: a Specification-Based Approach

David Sheeren[1,2], Sébastien Mustière[1], Jean-Daniel Zucker[3]

COGIT Laboratory - IGN France, 2-4 avenue Pasteur, 94165 Saint Mandé
{David.Sheeren,Sebastien.Mustiere}@ign.fr[1]
LIP6 Laboratory, AI Section, University of Paris 6[2]
LIM&BIO, University of Paris 13, Jean-Daniel.Zucker@lip6.fr[3]

Abstract

There currently exist many geographical databases that represent a same part of the world, each with its own levels of detail and points of view. The use and management of these databases therefore sometimes requires their integration into a single database. The main issue in this integration process is the ability to analyse and understand the differences among the multiple representations. These differences can of course be explained by the various specifications but can also be due to updates or errors during data capture. In this paper, we propose an new approach to interpret the differences in representation in a semiautomatic way. We consider the specifications of each database as the "knowledge" to evaluate the conformity of each representation. This information is grasped from existing documents but also from data, by means of machine learning tools. The management of this knowledge is enabled by a rule-based system. Application of this approach is illustrated with a case study from two IGN databases. It concerns the differences between the representations of traffic circles.

Keywords. Integration, Fusion, Multiple Representations, Interpretation, Expert-System, Machine Learning, Spatial Data Matching.

1. Introduction

In recent years, a new challenge has emerged from the growing availability of geographical information originating from different sources: their com-

bination in a consistent way in order to obtain more reliable, rich and useful information. This general problem of information fusion is encountered in different domains: signal and image processing, navigation and transportation, artificial intelligence… In a database context, this is traditionally called "integration" or "federation" [Sheth and Larson 1990, Parent and Spaccapietra 2000].

Integration of classical databases has already been given much attention in the database community [Rahm and Bernstein 2001]. In the field of geographical databases, this is also subject to active research. Contributions concern the integration process itself (the *schemas integration* and the *data integration*) [Devogele et al. 1998, Branki and Defude 1998], the development of matching tools [Devogele 1997, Walter and Fritsch 1999], the definition of new models supporting multiple representations [Vangenot et al. 2002, Bédard et al. 2002], and new data structures structures [Kidner and Jones 1994]. Some ontology-based approaches are now being proposed [Fonseca et al. 2002].

But some issues still need to be addressed in the process of unifying geographical databases, particularly the phase of *data integration*, i.e. the actual population of the unified database. Generally speaking, this phase is mainly thought of as a matching problem. However, it is also essential to determine if the differences in representation between homologuous objects are "normal", i.e. originating from the differences of specification.

Some contributions exist to evaluate the consistency between multi-representations, especially the consistency of spatial relations [Egenhofer et al. 1994, El-Geresy and Abdelmoty 1998, Paiva 1998]. Most of the time, studies rely on a presupposed order between representations. This assumption may not be adapted to the study of databases with similar levels of detail, but defined according to different points of view.

In this paper, we too address the issue of the assessment of consistency between multiple representations. The approach we suggest is based on the use of specifications from individual databases. We consider these specifications as the key point to understand the origin of the differences, and we suggest the use of a rule base to explicitly represent this knowledge and manage it. We make no assumptions on a hierarchy between the databases that need integrating.

The paper is organised as follows: in section 2, we examine the origins of the differences and the specification-based approach to interpret them. Then we propose an interpretation process and the architecture of the system to implement it in section 3. The feasibility of the approach is demonstrated with a particular application in section 4. We conclude our study in section 5.

2. Specifications for Interpreting Differences

2.1. Origin of Differences Between Representations

Geographical databases are described by means of specifications. These documents describe precisely the contents of a database, i.e. the meaning of each part of the data schema, which objects of the real world are captured, and how they are represented (figure 1).

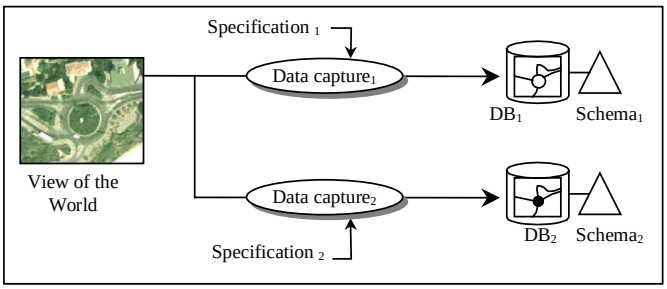

Fig. 1. Specifications govern the representation of geographical phenomena in databases

The differences between specifications are responsible for the majority of the differences between representations. These differences are completely normal and illustrate the diversity of points of views turned on the world. For example, a traffic circle may be represented by a dot in one database, or by a detailed surface in another. However, all differences are not justified. The data capture process is not free of errors and differences can occur between what is supposed to be in the databases and what is actually in the databases. Some other differences are due to the differences of update between databases. These differences are problematic because they can lead to inconsistent representations in the multi-representation system and for that reason, they must be detected and managed in a unification process.

More formally, we define hereafter the concepts of *equivalence*, *inconsistency* and *update* between representations. Let O be the set of objects from a spatial database DB_1 and O' the set of objects from a spatial database DB_2. Let us consider a matching pair of the form (M,M'), where M is a subset of O and M' a subset of O'.

Definition 1 (equivalence). *Representations of matching pairs (M,M') are said to be equivalent if these representations can model a world such as, at the same*

time, M and M' respect their specifications and correspond to the same entity of the real world.

Definition 2 (update). *Representations of matching pairs (M,M') are said to be of different periods if these representations can model a world such as M and M' respect their specifications and correspond to the same entity of the real world, but at different times.*

Definition 3 (inconsistency). *Representations of matching pairs (M,M') are said to be inconsistent if they are neither an update, nor an equivalence. Thus either M or M' does not respect its specifications (error in databases), or M and M' do not correspond to the same entity of the real world (matching error).*

The purpose of our work is to define a process to automatically detect and interpret differences between databases. This process, embedded in a decision support system, aims at guiding the management of differences during a unification process.

2.2. Knowledge Acquisition for the Interpretation of the Differences

The key idea of this approach is to make explicit, in an expert-system, the knowledge necessary to interpret the differences. As explained above, a great deal of the knowledge comes from the specifications. Nevertheless, it is rather difficult to draw knowledge from the specifications and to represent it. Actually, the documents are usually rich but voluminous, relatively informal, ambiguous, and not always organised in the same way. Moreover, part of the necessary knowledge comes from common geographical knowledge (for example: *traffic circles are more often retained than discarded*) and experts are rarely able to supply an explicit description of the knowledge they use in their reasoning. We are thus faced with the well-known problem of the "*knowledge acquisition bottleneck*". In our process, we try to solve it in three different ways.

The first technique is to split up into several steps the reasoning involved in the problem solving (section 3). This is the approach of second generation expert-systems [David et al 1993]. The control over the inferences that need to be drawn is considered as a kind of knowledge in itself, and explicitly introduced in expert-systems.

The second technique is to develop rules by hand with the help of a knowledge acquisition process. We believe that such a process should rely on the definition of a formal model of specifications [Mustière et al 2003]. In the example developed in section 4, some of the rules managed by the expert-system have been introduced by hand, after formalising the actual

specifications by means of a specific model. This is still ongoing research and it will not be detailed in this paper.

The last technique is the use of supervised machine learning techniques [Mitchell 1997]. These techniques are one of the solutions developed in the Artificial Intelligence field. Their aim is to automatically build some rules from a set of examples given by an expert. These rules can then be used to classify new examples introduced into the system. Such techniques have already been used to acquire knowledge in the geographical domain [Weibel et al 1995, Sester 2000, Mustière et al 2000, Sheeren 2003].

3. The Interpretation Process

3.1. Description of the Steps

In this section, we describe the interpretation process we have defined. It is decomposed in several steps which are illustrated in figure 2.

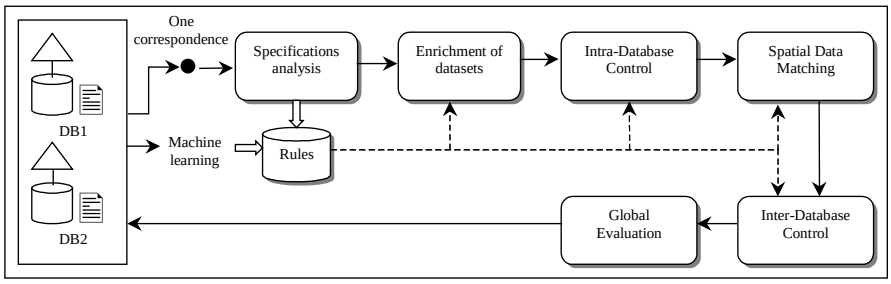

Fig. 2. From individual databases to the interpretation of differences

The process starts with *one correspondence* between classes of the schemas of the two databases. We presume that matching at the schema level has already been carried out. For instance, we know that the *road* class in DB1 tallies with the *road* and *track* classes in DB2.

The task of *specifications analysis* is then the key step: the specifications are analysed in order to determine several rule bases that will be used to guide each of the ensuing steps. These rules primarily describe what exactly the databases contain, what differences are likely to appear, and in which conditions. This step is performed through the analysis of documents, or through machine learning techniques.

The next step concerns the *enrichment* of each dataset. This is compulsory before the actual integration of the databases [Devogele et al 1998].

The purpose is to express the heterogeneous datasets in a more homogeneous way. For this step, the particularity of the geographical databases arises from the fact that they contain lots of implicit information on spatial relations through the geometry of objects. Their extraction requires specific analysis procedures.

A preliminary step of control is then planned: the *intra-database control*. During this step, part of the specifications is checked so as to detect some internal errors and determine how the data instances globally respect specifications. This will be useful for the identification of the origin of each difference but also, for the detection of matching errors.

Once the data of both databases has been independently controlled, it is matched. Matching relationships between datasets are computed through geometric and topologic data matching. We end up with a set of matching pairs, each one characterised by a degree of confidence.

The next step consists in the comparison of the representations of the homologous objects. This is the *inter-database control*. This comparison leads to the evaluation of the conformity of the differences and particularly implies the use of specifications and expert knowledge. Results of the first control previously carried out are also exploited. At the end, differences existing between each matching pair are expressed in terms of *equivalence*, *inconsistency* or *update.*

After the automatic interpretation of all the differences by means of the expert-system, a global evaluation is supplied: the number of equivalencies, the number of errors and their seriousness, and the number of updates.

3.2. The Architecture of the System

An illustration of the structure of the system is given in figure 3. It is composed of two main modules: the experimental *Oxygene GIS* and the *Jess* expert-system. *Oxygene* is a platform developed at the COGIT laboratory [Badard & Braun 2003]. Spatial data is stored in the relational *Oracle* DBMS, and the manipulation of data is performed with the *Java* code in the object-oriented paradigm. The mapping between the relational tables and the *Java* classes is done by the *OJB* library. A *Java* API exists to make the link between this platform and the second module, the *Jess* rule-based system. It is an open source environment which can be tightly coupled to a code written in *Java* language [Jess 2003]. The rules used by *Jess* originate directly from the specifications, or have been gathered with the learning tools.

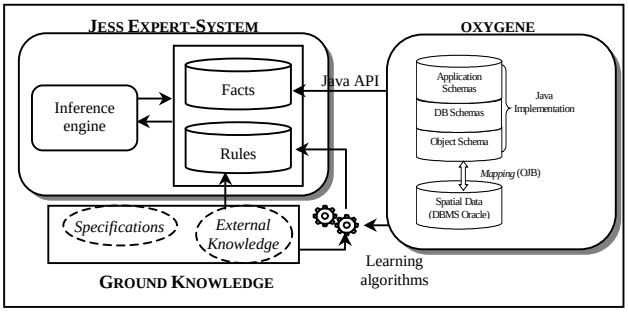

Fig. 3. Architecture of the system

4. Differences Between Representations of Traffic Circles: a Case Study

In this section, we study the differences existing between the traffic circles of two databases from the IGN (French National Mapping Agency): BDCarto and Georoute (figure 4). BDCarto is a geographical database meant in particular to produce maps at a scale ranging from 1:100,000 to 1:250,000. Georoute is a database with a resolution of 1 m dedicated to traffic applications. The representation of the traffic circles can differ from one database to another because the specifications are different. Our question is thus as follow: which differences are “normal”, i.e. which representations are equivalent, and which differences are “abnormal”, i.e. which representations are inconsistent ? We detailed the implementation of the process below.

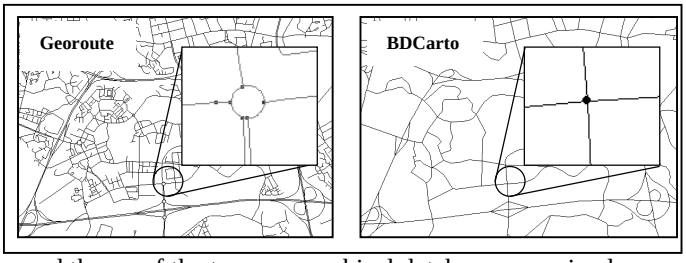

Fig. 4. The road theme of the two geographical databases examined

Specifications analysis. Specifications of BDCarto and Georoute explicitly describe the representation of the traffic circles. For both databases the representation can be simplified (corresponding to a node) or detailed (corresponding to connected edges and nodes). The modelisation depends

on the diameter of the object in the real world, but also on the presence of a central reservation (figure 5).

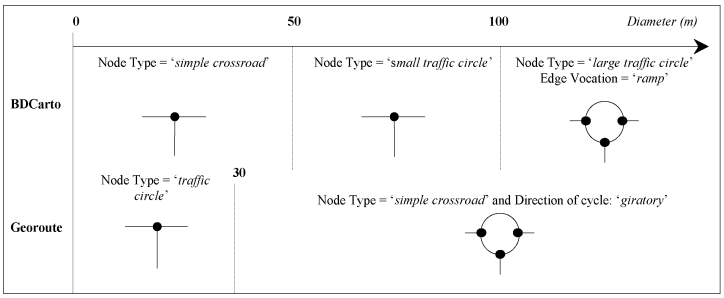

Fig. 5. Some specifications concerning traffic circles of BDCarto and Georoute

Specifications introduce differences between datasets. They appear both at geometry and attribute levels. The description also reveals a difficulty that has already been brought up: the gap existing between the data mentioned in the specifications and the data actually stored in the databases. The traffic circles are implicit objects made of several edges and nodes, but there is no corresponding class in the database. In the same way, the diameter of the objects and the direction of the cycle does not exist as an attribute in the databases. It is thus necessary to extract this information in order to check the specifications and enable the comparison between the data. This enrichment is the subject of the next step.

Enrichment of the data. In the unification context, the enrichment of the databases concerns both geometrical data and schemas. In figure 6, we illustrate the new classes and relations created at the schema level. These classes can constitute federative concepts to put in correspondence the two schemas of the databases during the phase of creating the unified schema [Gesbert 2002].

At the data level, it is also necessary to extract implicit information and instanciate the new classes and relations created. Several operations have been carried out to achieve this, for the two databases (figure 7). First, we created the *simple traffic circles* and their relation with the *road nodes.* The simple traffic circles are road nodes for which the attribute *nature* takes the value '*traffic circle*'. Concerning the *complex traffic circles*, the construction of a topological graph was first necessary. Faces were created and all the topological relations between edges, nodes and faces were computed. We then filtered each face in order to retain only those corresponding to a traffic circle. Several criteria were taken into account: the direction of the cycle, the number of nodes for each cycle and the value of Miller's circularity index. These criteria were embedded in rules and com-

bined with the decision support system. In doing so, only faces corresponding to a traffic circle were retained.

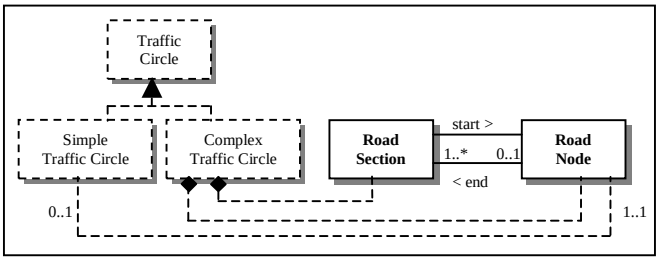

Fig. 6. Extract of the Georoute schema: new classes and relations are in dashed line.

The enrichment phase is thus performed to extract the implicit information required for the specifications control, but also to bring the structure of the data and schemas closer to each other.

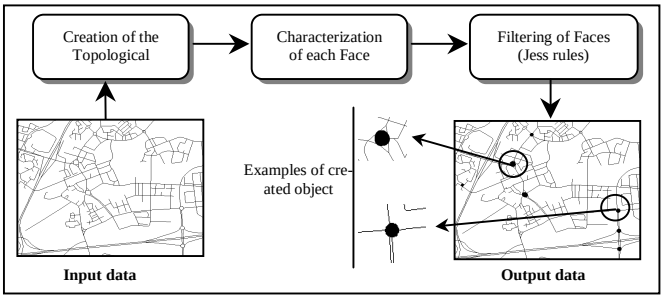

Fig. 7. The creation of complex traffic circles (extract of Georoute).

Intra-database Control. Two kinds of traffic circles were created in the previous step: simple traffic circles (nodes) and complex traffic circles (connected edges and nodes). At this level, the representations of the objects were checked to detect some internal errors. The control was automated thanks to several rules activated by the expert-system. These rules were developed and introduced by hand. For example:

```
(defrule control_diameter_georoute
(if diameterLength > 30)
      =>
(set diameterConformity "conform"))
```

Only part of the representations were controlled at that stage for each database: the complex traffic circles and the information associated with them (the diameter, the number of nodes,…). The node representation was checked later during the inter-database control because of the lack of in-

formation at that point. Some errors were identified during this process and the results were stored in specific classes for each database.

Spatial Data Matching. The matching tools used in this process are those proposed by [Devogele 1997]. They are founded on the use of both geometric and topologic criteria. They have been enriched by using the polygonal objects created during the previous steps, in order to improve the reliability of the algorithms. A degree of confidence was systematically given for each matching pair, according to the cardinality of the link, the dimension of the objects constituting the link and the matching criteria used. Finally, we have retained 89% of matching pairs, for a total of 690 correspondences computed.

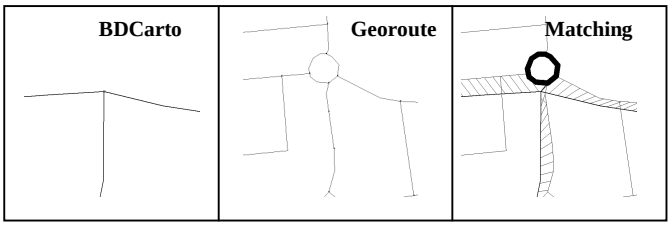

Fig 8. Example of homologous traffic circles and roads matched

Inter-database Control. Some internal errors were already detected during the first step of control but the representations of the two databases had not been compared. The comparison was the purpose of this step. It led to the classification of each matching pair in terms of *equivalence* and *inconsistency* (no updates were found for these datasets).

The introduction of rules by hand to compare the representations was first considered, but because of numerous possible cases and the complexity of some rules, we decided to use supervised machine learning. An example of a rule computed by the C5.0. algorithm [Quinlan 1994] is presented below. It enables the detection of an inconsistency:

If the type of the traffic circle in Georoute = 'dot'
And if the node type of the traffic circle in BDCarto = 'small traffic circle'
Then the representations are inconsistent

A set of rules have been introduced in the expert-system and finally, the set of matching pairs have been interpreted automatically. We have computed 67% of equivalencies and 33% of inconsistencies. Various types of inconsistencies were highlighted: modelling errors, attribute errors and geometrical errors (the variation between the diameters of the detailed objects were sometimes too high). We have noted that the errors were more frequent in the BDCarto.

5. Conclusion and Future Work

This paper has presented an new approach to deal with the differences in representation during the phase of *data integration* of geographical databases. The key idea of the approach is to use the specifications of each database to interpret the origin of the differences: *equivalence*, *inconsistency* or *updates*. The knowledge is embedded in rules and handled by an expert-system. The rules are introduced in two ways: by hand and thanks to supervised machine learning techniques.

This approach opens up many new prospects. It will be possible to improve the quality and up-to-dateness of each analysed database. The specifications could also be enriched and described in a more formal way. The use of the specifications and representations of one database can indeed help precise the capture constraints of the other database. Finally, we think that the study of the correspondences between the data could help find the mapping between the elements at the schema level. Few research have been made in that direction.

6. References

Badard T. and Braun A. 2003. OXYGENE : an open framework for the deployment of geographic web services, *In Proceedings of the International Cartographic Conference*, Durban, South Africa, pp. 994-1003.

Bédard Y., Bernier E. et Devillers R. 2002. La métastructure vuel et la gestion des représentations multiples. In *Généralisation et représentation multiple*, A. Ruas (ed.), chapitre 8.

Branki T. and Defude B. 1998. Data and Metadata: two-dimensional integration of heterogeneous spatial databases, In *Proceedings of the 8th International Symposium on Spatial Data Handling*, Vancouver, Canada, pp. 172-179.

David J.-M., Krivine J.-P. and Simmons R. (eds.) 1993. Second Generation Expert Systems, Springer Verlag.

Devogele T. 1997. Processus d'intégration et d'appariement de bases de données Géographiques. Application à une base de données routières multi-échelles, PhD Thesis, University of Versailles, 205 p.

Devogele T., Parent C. and Spaccapietra S. 1998. On spatial database integration, *International Journal of Geographical Information Science*, 12(4), pp.335-352.

Egenhofer M.J., Clementini E. and Di Felice P. 1994. Evaluating inconsistencies among multiple representations, In *Proceedings of the Sixth International Symposium on Spatial Data Handling*, Edinburgh, Scotland, pp. 901-920.

El-Geresy B.A. and Abdelmoty A.I. 1998. A Qualitative Approach to Integration in Spatial Databases, In *Proceedings of the 9th International Conference on Database and Expert Systems Applications*, LNCS n°1460, pp. 280-289.

Fonseca F.T., Egenhofer M., Agouris P. and Câmara G. 2002. Using ontologies for integrated Geographic Information Systems, *Transactions in GIS*, **6**(3).

Gesbert N. 2002. Recherche de concepts fédérateurs dans les bases de données géographiques, *Actes des 6ème Journées Cassini*, École Navale, pp. 365-368.

Jess 2003. The Jess Expert-System, http://herzberg.ca.sandia.gov/jess/

Kidner D.B. and Jones C. B. 1994. A Deductive Object-Oriented GIS for Handling Multiple Representations, In *Proceedings of the 6th International Symposium on Spatial Data Handling*, Edinburgh, Scotland, pp. 882-900.

Mitchell T.M. 1997. Machine Learning. McGraw-Hill Int. Editions, Singapour.

Mustière S., Zucker J.-D. and Saitta L. 2000. An Abstraction-Based Machine Learning Approach to Cartographic Generalisation, In *Proceedings of the 9th International Symposium on Spatial Data Handling*, Beijing, pp. 50-63.

Mustière S., Gesbert N. and Sheeren D. 2003. A formal model for the specifications of geographic databases, In *Proceedings of the 2nd Workshop on Semantic Processing of Spatial Data* (GeoPro'2003), Mexico City, pp. 152-159.

Paiva J.A. 1998. *Topological equivalence and similarity in multi-representation geographic databases*, PhD Thesis, University of Maine, 188 p.

Parent C. and Spaccapietra S. 2000. Database Integration: the Key to Data Interoperability. In *Advances in Object-Oriented Data Modeling*, Papazoglou M., Spaccapietra S. and Tari Z. (eds). The MIT Press.

Quinlan J.R. 1993. C4.5 : Programs for machine learning, Morgan Kaufmann.

Rahm E. and Bernstein P.A. 2001. A survey of approaches to automatic schema matching, *Very Large Database Journal*, 10, pp. 334-350.

Sester M. 2000. Knowledge Acquisition for the Automatic Interpretation of Spatial Data, *International Journal of Geographical Information Science*, 14(1), pp. 1-24.

Sheeren D. 2003. Spatial databases integration : interpretation of multiple representations by using machine learning techniques, In *Proceedings of the International Cartographic Conference*, Durban, South Africa, pp. 235-245.

Sheth A. and Larson J. 1990. Federated database systems for managing distributed, heterogeneous and autonomous databases, *ACM Computing Surveys*, 22(3), pp. 183-236.

Vangenot C., Parent C. and Spaccapietra S. 2002. Modeling and manipulating multiple representations of spatial data, In *Proceedings of the International Symposium on Spatial Data Handling*, Ottawa, Canada, pp. 81-93.

Walter V. and Fritsch D. 1999. Matching Spatial Data Sets: a Statistical Approach, *International Journal of Geographical Information Science*, 13(5), pp. 445-473.

Weibel R., Keller S. et Reichenbacher T. 1995. Overcoming the Knowledge Acquisition Bottleneck in Map Generalization : the Role of Interactive Systems and Computational Intelligence, In *Proceedings of the 2nd International Conference on Spatial Information Theory*, pp. 139-156.