Compact, evolving community taxonomies using concept lattices

Camille Roth

CREA (Center for Research in Applied Epistemology), CNRS/Ecole Polytechnique, Paris, France *and* Dept. of Social and Cognitive Science, University of Modena & Reggio Emilia, Italy
camille.roth@polytechnique.edu

Abstract. We introduce a formal framework based on formal concept lattices, or Galois lattices, that categorizes epistemic communities automatically and hierarchically, rebuilding a relevant taxonomy in the form of a *hypergraph of significant epistemic sub-communities*. The longitudinal study of these static pictures makes historical description possible, by capturing epistemological descriptions through stylized facts such as field emergence, decline, specialization and interaction (merging or splitting). The method is applied to empirical data describing the evolution of a particular community of scientists (embryologists working on the model-animal "zebrafish"), and successfully validated by categories and *histories* given by domain experts.

Keywords: Concept lattices / Galois lattices, Applied epistemology, Dynamic taxonomies, Knowledge discovery in databases, Scientometrics, Categorization.

Introduction

Scientists, journalists, socio-cultural communities with common references are various instances of the society of knowledge, being smaller, embedded societies with specific topics; partially independent, partially overlapping. Any knowledge community — the whole society, biologists, embryologists working on a particular model-animal — appears to be structured in turn in various subcommunities contributing to knowledge creation in a decentralized and complementary manner. Expertise is indeed heterogenously distributed over all agents: boundaries appear between subgroups, both horizontally, with distinct areas of competence, and vertically, with different levels of generality for appraising a given topic.

Yet, while agents can potentially access a large part of the knowledge produced by the whole epistemic community, they actually know only a small portion of it, prominantly because of cognitive and physical limitations. In this respect, it is of utmost interest to have tools enabling agents to understand the activity of their knowledge community, at any level of specificity or generality. More precisely, agents have an implicit (meta-)knowledge of the structure of the larger global community they are participating in: embryologists know what molecular biology, biology, and science in general are about. But this knowledge, limited and subjective, resembles that of a folk taxonomy, in the anthropological

sense [28], that is, a taxonomy proper to an individual and made of its own experience, as opposed to scientific taxonomies, deemed objective and systematic [2]. Hence, epistemologists often have the last word in elaborating and validating credible meta-knowledge: expert-made taxonomies are prodigiously more reliable than folk taxonomies, but still lack precision, take an enormous amount of work, and rarely investigate comprehensively the whole community.

We use *formal concept* lattices, or Galois lattices, to appraise the taxonomy of a given knowledge community, by building and ordering its *epistemic hypergraph*. An epistemic hypergraph is a graph of *epistemic communities*, gathering both agents and notions (or concepts, in the usual sense). Essentially, an epistemic community is a group of agents interested in some common knowledge issues, using shared paradigms and meanings [9, 17]: for instance a group of research investigating a precise topic. This is however not necessarily restricted to academic groups, and an epistemic community needs not be a community of practice either [26]; although a community of practice is certainly a special type of knowledge community.

The produced epistemic hypergraph must make clear (i) which fields & trends are to be found, (ii) what kind of relationships they entertain. In turn, the resulting taxonomy should prove consistent with the already-existing intersubjective perception of the field, which will thus be the benchmark of our procedure. Eventually, knowing the taxonomy at any given time enables us to describe the evolution of the system; and as such achieve a reconstruction of the history of the community *on objective grounds*.

The outline of the paper is as follows: after having presented the context and introduced the formal framework (Sec. 1), we describe how to categorize epistemic communities in an hierarchically structured fashion using concept (Galois) lattices (Sec. 2) and produce a manageable (reduced) lattice-based representation of the whole knowledge community. We also adress their evolution: in particular, field progress or decline, field scope enrichment or impoverishment, and field interaction (merging or splitting) are defined. The method is eventually applied to a dynamic case study (Sec. 3). Settled both in applied epistemology and scientometrics, this approach would ultimately provide agents with processes enabling them to know dynamically their community structure.

1 Formal framework

1.1 Context

For social epistemologists, an *epistemic community* is a group of scientists producing knowledge and recognizing a given set of conceptual tools and representations [9, 17] — the "paradigm," according to Kuhn [24]. Several formal frameworks and automated processes have been proposed to analyze knowledge community structure and find groups of agents or documents related by common notions or concerns, notably in knowledge discovery in databases (KDD) [20] and scientometrics [6, 27], following the development of massive informational content (in particular scientific data). Yet, most approaches in community finding

are either based on social relationships only, with community extraction methods stemming from graph theory applied to social networks [37], or on semantic similarity only, namely clustering methods applied to document databases where each document is considered as a vector in a semantic space [33]. There have been few attempts to link social and semantic aspects, although the various characterizations of an epistemic community insist on its duality: a community is on one side a group of agents who, on the other side, work on a given subset of notions.

On the other hand, many different techniques have been proposed for producing categorical structures including, to cite a few, hierarchical clustering [21], *Q*-analysis [1], formal concept analysis [40], blockmodeling [39], graph theory-based techniques [31], neural networks [23]. Along with this profusion of community-finding methods, often leaning towards AI-oriented clustering, an interesting issue concerns the representation of communities in an ordered fashion. Here, the notion of *taxonomy* is particularly relevant with respect to communities of knowledge. Taxonomies are hierarchical structurations of categories (or ordered set of *taxons*), useful in biology, cognitive psychology, as well as ethnography and anthropology; and while they have initially been built using a subjective approach, the focus has moved to formal and statistical methods [34].

However, taxonomy building itself is generally poorly investigated; arguably, taxonomy evolution during time has been fairly neglected. Our intent here is thus to address both topics: build a taxonomy of epistemic communities, then monitor its evolution. At the same time, while taxonomies have long been represented using tree-based structures, we wish to deal with sub-communities affiliated with multiple communities; thus calling for lattice-based structures.

1.2 Epistemic communities and epistemic hypergraphs: definitions

Basically, we try to know (i) which agents share the same concerns and work on the same notions, and (ii) which these concerns or notions are. Hence, our definition of an epistemic community (EC) is simply characterized by common knowledge concerns and should not necessarily be a social community:

Definition 1 (Epistemic community). *Given an agent set S, the* epistemic community *of S is the* largest *set of agents who use the notions which all agents of S have in common.*

Considering the epistemic community of an agent set extends it to the largest community sharing all its notions. This concept is close to "*structural equivalence*," introduced in sociology by F. Lorrain and H. White [29]: ECs are groups of agents related in an *equivalent* manner to some notions. We could also define correspondingly an epistemic community as the largest set of notions commonly used by agents who share a given notion set. We will at first focus on agent-based ECs, keeping in mind that notion-based concepts are defined strictly equivalently and in a dual manner.

Formally, we bind agents to notions with a binary relation $\mathcal{R}$ between the whole agent set $\mathbf{S}$ and the whole notion set $\mathbf{N}$. Here, $\mathcal{R} \subseteq \mathbf{S} \times \mathbf{N}$ represents any

kind of link between an agent s and a notion n: in our case, the link corresponds to the fact that s used n (e.g. in some article). Then, we define the "intent" of an agent set S as the the set of notions used by every agent in S — it is the set of elements of $\mathbf{N}$ $\mathcal{R}$-related to *every* element of S. Similarly and dually, we define the "extent" of a notion set N as the set of agents who use every notion in N. We denote the intent of S and extent of N by $S^\wedge$ and $N^\star$ respectively, thereby implicitly defining two operators:

Definition 2 (Intent and extent operators). *The intent operator "$\wedge$" is such that,* $\forall s \in \mathbf{S}$, $\forall S \subseteq \mathbf{S}$,

$$\begin{cases} s^\wedge = \{\, n \in \mathbf{N} \mid s\mathcal{R}n \,\} \\ S^\wedge = \{\, n \in \mathbf{N} \mid \forall s \in S, s\mathcal{R}n \,\} \end{cases} \tag{1}$$

and the extent operator "$\star$" is such that $\forall n \in \mathbf{N}$, $\forall N \subseteq \mathbf{N}$,

$$\begin{cases} n^\star = \{\, s \in \mathbf{S} \mid s\mathcal{R}n \,\} \\ N^\star = \{\, s \in \mathbf{S} \mid \forall n \in N, s\mathcal{R}n \,\} \end{cases} \tag{2}$$

By definition, we set $(\emptyset)^\wedge = \mathbf{N}$ *and* $(\emptyset)^\star = \mathbf{S}$.

This formalism is traditional in Formal Concept Analysis (FCA) [41] and is a robust way of dealing with abstract notions in a philosophical sense, characterized by their physical implementation (extent) and their internal content (intent). Here, notions are *properties* of authors who use them (they are skills in scientific fields, i.e. cognitive properties) and authors are *loci* of notions (notions are implemented in authors).

Furthermore, the combination of "$\wedge$" and "$\star$" yields a *closure operator*: for a given S, $S^{\wedge\star}$ is called the closure of S; we can say that "$\wedge\star$" is:

(i) extensive: $S \subseteq S^{\wedge\star}$ — the closure is never smaller;
(ii) idempotent ($(S^{\wedge\star})^{\wedge\star} = S^{\wedge\star}$) — applying $\wedge\star$ once or more does not change the closure;
(iii) increasing ($S \subseteq S' \Rightarrow S^{\wedge\star} \subseteq S'^{\wedge\star}$) — the closure of a larger set is larger.

Thus, applying "$\wedge\star$" to S returns *all the agents* who use the same notions which agents of S had in common. Briefly, $\wedge\star$ yields the EC of any agent set: ($S^{\wedge\star}$, $S^\wedge$) is the epistemic community based on S.

Definition 3 (Closed couple). *Given two subsets* $S \subseteq \mathbf{S}$ *and* $N \subseteq \mathbf{N}$, *a couple* (S, N) *is said to be* closed *if and only if* $N = S^\wedge$ *and* $S = N^\star$.

Subsequently, any closed couple is an epistemic community, which we can denote unambiguously and indifferently by its agent set or its notion set. In FCA, such a couple is classically called a "formal concept." We may now introduce *epistemic hypergraphs*:

Definition 4 (Hypergraph, epistemic hypergraph). *A hypergraph* $\mathbf{H}$ *is a couple* (V, E) *where* V *is a set of vertices and* E *a set of hyperedges connecting a set of vertices.* E *is thus fundamentally a subset of* $\mathcal{P}(V)$, *the power set of* V. *An epistemic hypergraph is a hypergraph of ECs,* $(\mathbf{S}, \{S^{\wedge\star} | S \subset \mathbf{S}\})$ *with hyperedges binding groups of agents belonging to a same EC.*

Each hyperedge can be labelled with the notion set corresponding to the agent set it binds, $S^\wedge$. An epistemic hypergraph is basically the set of all ECs.[1]

2 Galois lattices: from relations to dynamic taxonomies

2.1 Taxonomies, lattices and epistemic hypergraphs

A relationship between agents and notions is thus sufficient to capture the whole underlying epistemic hypergraph of a given scientific field, and as such it is yet another clustering method. Is it also able to capture a *meaningful* structure? There are several stylized facts we would like to rebuild, primarily the existence of subfields and significant groups of agents working within those subfields. One can consider epistemic hypergraphs from any two sets of objects and a given relationship between them, yet there is no reason *a priori* why this should reveal a remarkable structure.

Our main assumption is that there are fields of knowledge which can be described by notion lists, and which are being implemented by sets of agents. For instance, some scientists are linguists, and some among them deal with a given aspect, say prosody; some other scientists deal with neuroscience, while a few of them are interdisciplinary and use both notions. Moreover, these fields are hierarchically organized: a general field can be divided into many subfields, themselves possibly having subcategories or belonging to various general fields, being *multi-disciplinary* or *inter-disciplinary* in that they respectively involve or integrate two or more subfields [22].

To hierarchize the raw set of all ECs that makes the epistemic hypergraph, we first provide a *partial order* between ECs:

Definition 5 (Subfield). *An EC* $(S, S^\wedge)$ *is a* subfield *of the EC* $(S', S'^\wedge)$ *if its intent is larger (more precise), or equivalently if its extent is smaller:*

$$(S, S^\wedge) \sqsubset (S', S'^\wedge) \Leftrightarrow S \subset S' \tag{3}$$

We can thus render both generalization and specification of closed couples [41], because $(S, S^\wedge)$ can be seen as a specification of $(S', S'^\wedge)$ (larger notion set, less agents) and conversely $(S', S'^\wedge)$ is a *"superfield"* or a generalization of $(S, S^\wedge)$. Now, the *concept lattice*, or *Galois lattice* [3] is exactly the ordered set of all epistemic communities built from $\mathbf{S}$, $\mathbf{N}$ and $\mathcal{R}$:

Definition 6 (Galois lattice). *The* Galois lattice $\mathcal{G}_{\mathbf{S},\mathbf{N},\mathcal{R}}$ *is the set of every* closed *couple* $(S, N) \subseteq \mathbf{S} \times \mathbf{N}$ *under relation* $\mathcal{R}$*:* $\mathcal{G}_{\mathbf{S},\mathbf{N},\mathcal{R}} = \{(S^{\wedge\star}, S^\wedge) | S \subseteq \mathbf{S}\}$*, partially ordered with* $\sqsubset$*.*

[1] Note that all these properties are dual when considering an EC based on N, subset of $\mathbf{N}$, and "$\star\wedge$" — for instance, an epistemic hypergraph could equivalently be based on notions: $(\mathbf{N}, \{N^{\star\wedge} | N \subset \mathbf{N}\})$, with hyperedges binding notions of a same EC.

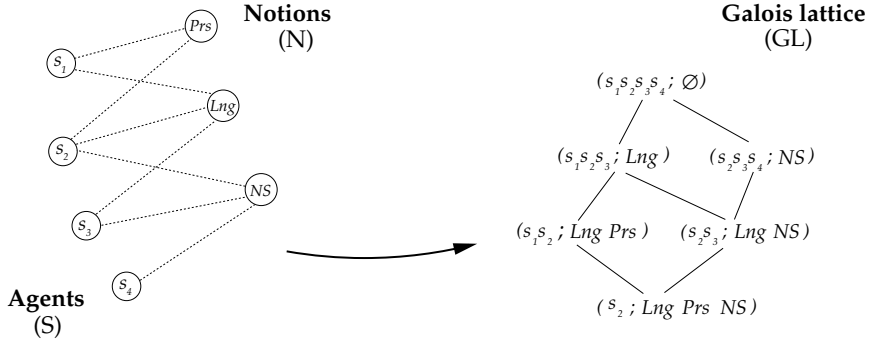


Fig. 1. Creating a Galois lattice of 6 ECs (left) from a sample community (right), involving agents s_1, s_2, s_3, s_4 and notions *"linguistics"* (Lng), *"neuroscience"* (NS) and *"prosody"* (Prs). ECs are a pair (extent, intent) $= (S, N)$ with $S^{\wedge} = N$, $N^{\star} = S$. An EC closer to the top is more general: the hierarchy reproduces the generalization/specialization relationship induced by $\sqsubset$.

Relevance of Galois lattices. Lattices in general replace efficiently and conveniently trees for describing such taxonomies — trees are the canonical (Aristotelian) approach for ordering categories, where sub-categories are child nodes of their unique parent category, thus unable to deal with category overlap, or weak at representing paradigmatic categories. Using GLs we represent ECs hierarchically in a lattice-based taxonomy. More broadly, GLs are suitable for ordering abstract categories relying on such a binary relation, and have been therefore widely used in conceptual knowledge systems, formal concept classification, as well as mathematical social science [12, 14, 15, 30, 40]. GLs can also be considered as hierarchically ordered epistemic hypergraphs — as such, GLs are both a categorization tool and a taxonomy building method. A graphical representation of a GL is drawn on Fig. 1.

GL relevance for our purpose results from the fact that (i) knowledge fields and their corresponding agent sets are ECs, which are precisely what GLs consist of, (ii) the GL natural partial order $\sqsubseteq$ reflects a generalization/specialization relationship between fields and subfields and exhibits multidisciplinarity and interdisciplinarity. Assuming this organization of scientific communities, the justification for this method will lie in the agreement between EC taxonomies extracted using GLs and those explicitly given by domain experts.

2.2 Trimming the lattice

However, a serious caveat of GLs is that they may grow extremely large, with significantly more than several thousands of ECs, even for few agents and notions. GLs contain all ECs, and among those many do not correspond to an existing or relevant field of knowledge: how to produce a *useful* and *usable* representation? We should select *relevant ECs* from a possibly huge GL, while

excluding irrelevant ones. Formally, the new epistemic hypergraph that contains only extracted ECs is still a partially-ordered set (with $\sqsubseteq$), which overlays on the lattice structure and enjoys the taxonomical properties we are interested in. The resulting concise taxonomical description is called hereafter "*partial epistemic hypergraph.*" This selection process has so far been an underestimated topic in the study of GLs: an important part of the effort has focused on computation and representation [10, 15, 25], while few authors insist on the need for semantic interpretations and approximation theories in order to cope with GL combinatorial complexity [12, 35, 36].

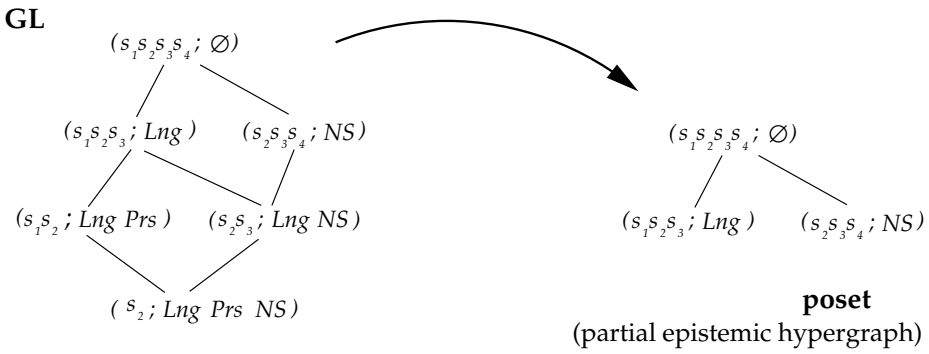


Fig. 2. From the original GL to a selected poset, or *partial epistemic hypergraph.*

Criteria. We need to design criteria for distinguishing relevant ECs. At first, we would certainly keep largest ECs: if a set of notions corresponds to a field, its extent should be of significant size [32]. Yet, some large ECs are too specific, while small ECs close to the top are also likely to be relevant, as minority fields. We thus suggest some selection properties: (i) agent set size, (ii) level (shortest distance to the top), (iii) specificity (notion set size), (iv) sub-communities (number of descendants). Then, we design several simple *selection heuristics* attributing a score to each EC by combining these criteria, so that we only keep the top scoring ECs: for instance, favoring (i) large ECs, (ii) close to the top, (iii) unusually specific (i.e. featuring many notions with respect to what the position in the lattice would suggest), (iv) close to the top *and* having few descendants (possibily being heterodox ECs). We may not necessarily be able to express all preferences through a unique heuristic. Therefore, the selection process involves several heuristics: for instance one function could select large communities, while another is best suited for minority communities. Fine tuning these heuristics eventually requires active feedback from empirical data; yet in any case, correct results with respect to the rebuilding task acknowledges the validity of the choice.

Dealing with computational complexity. Practically, this pruning process can be achieved in two ways: either (i) by computing the whole GL then select a partial epistemic hypergraph, or (ii) by computing only relevant ECs. The former solution is certainly the least efficient option; while the latter addresses the theoretically exponential complexity by safely limiting the number of ECs to be computed. Yet, doing so requires an incremental construction of the lattice, which actually constrains strongly the choice of selection criteria. For instance, one could suggest computing the upper part and its "valuable" descendance — computing a fixed number of ECs, starting from the top, similarly to what "iceberg lattices" achieve [35] — but this requires monotonic selection heuristics, i.e. heuristics h respecting the lattice partial order: if $(S, N) \sqsubset (S', N')$, then $h(S, N) < h(S', N')$.

On the other hand, we are interested in scientific field taxonomy rather than monitoring a fixed and particular set of researchers. Therefore, considering a statistically significant random sample of authors should yield a partial epistemic hypergraph which faithfully accounts for the original thematic taxonomy. Here, computing the whole lattice then prune it is a satisfying solution, as long as the agent set is kept to a decent size. Besides, it is later possible to fill the surviving ECs with authors who were initially excluded from the computation.

2.3 Taxonomy evolution

We would also like to be able to provide an history of the field that matches an expert-based history, i.e. monitoring taxonomy evolution through partial epistemic hypergraph evolution, in a longitudinal study. This can be done by capturing some *patterns* reflecting epistemic evolution: (i) *progress or decline of a field*, (ii) *enrichment or impoverishment of a field* (reduction or extension of notion set related to a field), and (iii) *reunion or scission of fields* (emergence or disappearance of joint ECs made of several fields). In terms of changes between successive partial epistemic hypergraphs, these patterns simply translate into a variation in the population of a given EC. The interpretation of this population change ultimately depends on the EC position in the partial epistemic hypergraph: according to whether (i) the change concerns a single EC, (ii) it occurs for a subfield and (iii) this subfield is in fact a joint subfield. These patterns describe epistemic evolution with an increasing precision — Fig. 3. More precise patterns could naturally be proposed, yet these ones are sufficiently relevant for our purpose.

3 Case study

3.1 Empirical protocol

We now apply this procedure to an empirical case study: we considered the community of embryologists working on the model animal *"zebrafish"*, on the period 1990–2003, covering what experts of the field call the major growth of this

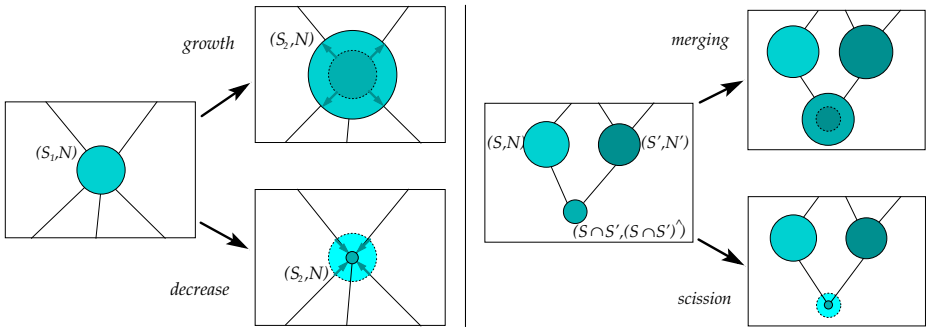


Fig. 3. *Left:* progress or decline of a given EC (S_1, N), whose agent set is growing (above) or decreasing (below). *Right:* emergence or disappearance of a joint community based on two ECs, (S, N) and (S', N'). Disk radius represents agent set size.

community, up to recent times. To describe the community evolution over several periods of time, we use data telling us *when* an agent s uses a notion n. Our main source of data is `MedLine`, a public bibliographical reference database maintained by the US National Library of Medicine. We adopt weak linguistic assumptions by assuming that a lemmatized term corresponds to a notion (lemmatization consists in finding the root of a word, or term: for instance, "*gener*" for both "*generic*" and "*general*"). We also restrict the dictionary to the 70 most used and significant words in the community, selected with the help of our expert in order to avoid rhetorical and neutral terms ("stop-words"). We attribute a notion to an agent whenever a lemmatized word is found in the title or the abstract of an article authored by the given agent.[2]

We divide the database into several time-slices, and build a series of relation matrices aggregating all events for each period. Before doing so, we choose the *time-slice width* (size of a period) and the *time-step* (increment of time between two periods):

1. *Time-slice width* — We need a sufficiently wide time-slice to take into account minority communities, to get enough information for each author, and smooth the data by reducing singularities. Yet we shall not merge several periods of evolution into a single time-slice: this tradeoff must be empirically

[2] This simplistic linguistic procedure could be easily improved by understanding terms in their context — for example, distinguishing several meanings for "*pattern*" as "*pattern-1*", "*pattern-2*", etc. Moreover, one could take into account simple semantic relationships such as synonymy and hyperonymy/hyponymy. Synonymy could be addressed by grouping several synonymic words under the same notion. Hyperonymic relationships could be rendered by simple implications: if n is an hyperonym of ν, then one could consider for instance that n should be added every time ν appears ($\nu \rightarrow n$). In the lattice, ECs using ν would necessarily be subfields of ECs using n.

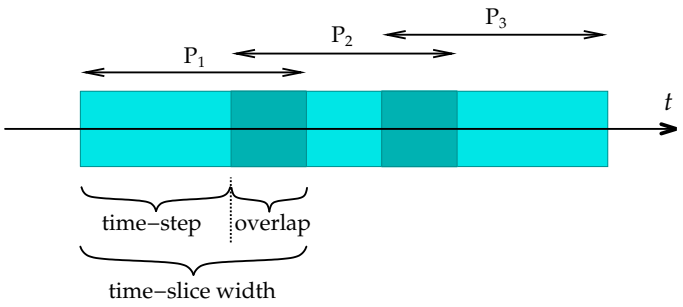


Fig. 4. Series of overlapping periods P_1, P_2 and P_3.

adapted to the data, for instance by choosing to talk in terms of months, years or decades.

2. *Time-step* — The time-step defines the pace of observation. Overlapping time-slices are needed to catch developments covering the end of a period and the beginning of the next one, so we have to choose a time-step strictly shorter than the time-slice width.

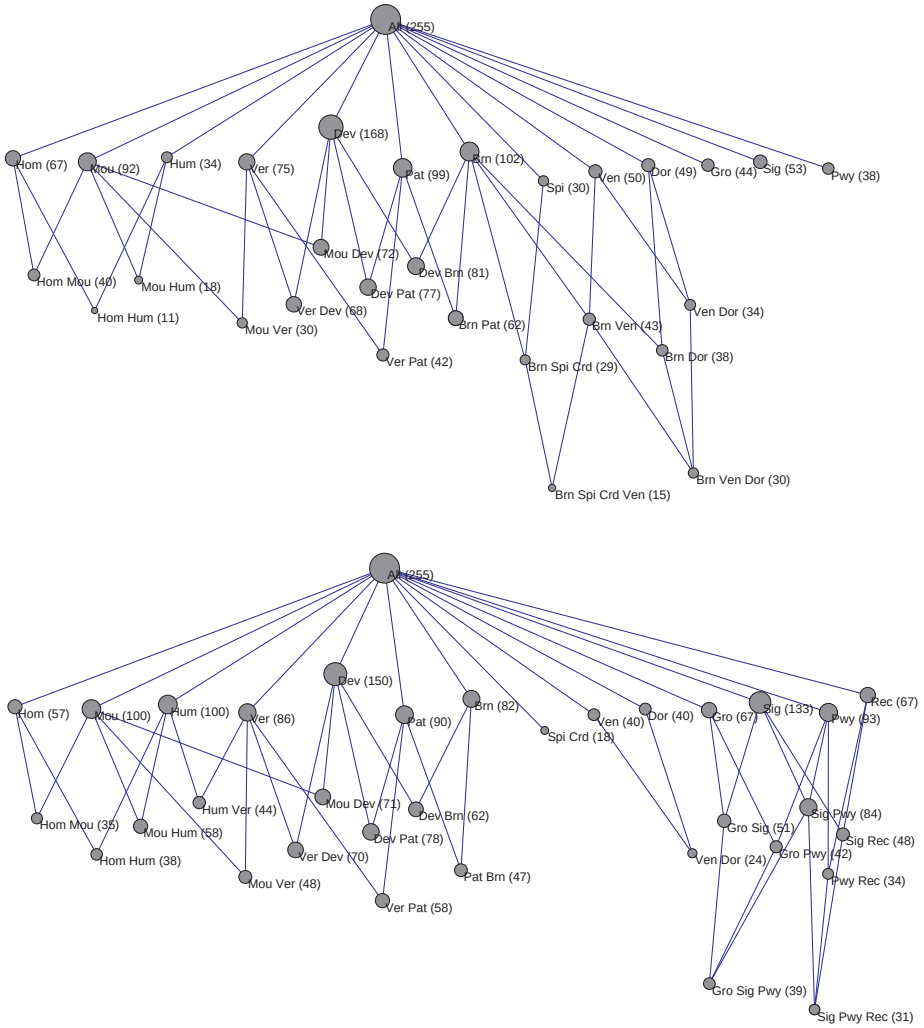
We splitted the database in three periods: 1990-1995, 1994-1999 and 1998-2003, i.e. a time-slice width of 6 years, with a time-step of 4 years — see Fig. 4. To limit computation costs, we also considered *for each period* a random sample of 255 authors. Besides with a fixed number of authors we could compare the relative importance of each field with respect to others within the evolving taxonomy.

3.2 Rebuilding history

Few changes occured between the first and the second period, and between the second and the third period: the second period is a transitory period between the two extreme periods. This seems to indicate that a 4-year time-step is slightly below the time-scale of the community, while 8 years can be considered a more significant time-scale. We hence focus on two periods: the first one, 1990-1995, and the third one, 1998-2003. The two corresponding partial epistemic hypergraphs are drawn on Fig. 5:

– *First period (1990-1995), first partial epistemic hypergraph:* {*develop*} and {*pattern*} strongly structure the field: they are both large communities and present in many subfields. Then, slightly to the right of the partial hypergraph, a large field is structured around *brain*[3] and *ventral* along with *dorsal*. Excepting one agent, the terms *spinal* and *cord* form a community with *brain*; this dependance suggests that the EC {*spinal*, *cord*} is necessarily linked to the study of *brain*. Subfields of {*brain*} also involve *ventral* and *dorsal*. Similarly, {*brain, ventral*} has a common subfield with {*spinal, cord*}.

[3] We actually grouped *brain*, *nerve*, *neural* and *neuron* under this term.



Legend: All: the whole community, Hom: *homologue/homologous,* Mou: *mouse,* Hum: *human,* Ver: *vertebrate,* Dev: *development,* Pat: *pattern,* Brn: *brain/neural/nervous/neuron,* Spi: *spinal,* Crd: *cord,* Ven: *ventral,* Dor: *dorsal,* Gro: *growth,* Sig: *signal,* Pwy: *pathway,* Rec: *receptor.*

Fig. 5. Two partial epistemic hypergraphs for period 1990–1995 (*top*) and 1998–2003 (*bottom*). Figures in parentheses indicate the number of agents per EC. Lattices established from a sample of 255 agents (out of 1,000 for the first period vs. 9,700 for the third one).

To the left, another set of ECs is structured around {*homologous*}, {*mouse*} and {*vertebrate*}, and {*human*}.

– *Third period (1998-2003), second partial epistemic hypergraph:* We still observe a strong structuration around {*develop*} and {*pattern*}, suggesting that the core topics of the field did not evolve. However, we notice the strong emergence of three communities, {*signal*}, {*pathway*} and {*growth*}, along with a new EC, {*receptor*}. These ECs form many joint subcommunities together, indicating a convergence of interests.
Also, there is a slight decrease of {*brain*}. More interestingly, there is no joint community anymore with {*ventral*} nor {*dorsal*}. The interest in {*spinal cord*} has decreased too, in a larger proportion. Finally, {*human*} has grown a lot, not {*mouse*}. These two communities are both linked to {*homologous*} on one side, {*vertebrate*} on the other. While the importance of {*homologous*} is roughly the same, the joint EC with {*human*} has increased a lot. The same goes with {*vertebrate*}: this EC, which is almost stable in size, has a significantly increased role with {*mouse*} and especially {*human*} (a new EC {*vertebrate, human*} just appeared).

Inference of an history. We summarize in terms of dynamic patterns (Sec. 2.3): some communities were stable (e.g. {*pattern*}, {*develop*}, {*vertebrate, develop*}, {*homologous, mouse*}), some enjoyed a burst of interest ({*growth*}, {*signal*}, {*pathway*}, {*receptor*}, {*human*}) or suffered less interest ({*brain*} and {*spinal cord*}). Also, some ECs merged ({*signal*}, {*pathway*}, {*receptor*} and {*growth*} altogether; and {*human*} both with {*vertebrate*} and {*homologous*}), some splitted ({*ventral-dorsal*} separated from {*brain*}). We did not see any *strict* enrichment or impoverishment — even if merging and splitting can be interpreted as such.

We can consequently suggest the following story:

(i) research on brain and spinal cord depreciated, weakened their link with ventral/dorsal aspects (in particular the relationship between ventral aspects and the spinal cord);
(ii) the community started to enquire relationships between signal, pathway, and receptors (all actually related to biochemical messaging), together with growth (suggesting a messaging oriented towards growth processes), indicating new very interrelated notions prototypical of an emerging field; and finally
(iii) while mouse-related research is stable, there has been a significant stress on human-related topics, together with a new relationship to the study of homologous genes and vertebrates, underlining the increasing role of {*human*} in these differential studies and their growing focus on human-zebrafish comparisons (leading to a new "interdisciplinary" field).

Point (ii) entails more than the mere emergence of numerous joint subcommunities: most pairs of notions in the set {*growth, pathway, receptor, signal*} are involved in a joint subfield. Put differently these notions form a clique of joint communities, a pattern which may be interpreted as *paradigm emergence* (see Fig. 5–bottom).

Comparison with real taxonomies. We compared these findings with empirical taxonomical data, coming both from:

1. Expert feedback: Our expert, Nadine Peyriéras, confirms that points (i), (ii) and (iii) in the previous paragraph are an accurate description of the field evolution. For instance, according to her, the human genome sequencing in the early 2000s [19] opened the path to zebrafish genome sequencing, which made possible a systematic comparison between zebrafish and humans, and consequently led to the development described in point (iii). In addition, the existence of a subcommunity with *brain*, *spinal cord* and *ventral* but not *dorsal* reminded her the initial curiosity around the ventral aspects of the spinal cord study, due to the linking of the ventral spinal cord to the mesoderm (notochord), i.e. the rest of the body.
2. Litterature: The only article comprehensively dealing with the history of this field seems to be that of Grunwald & Eisen [16]. This paper presents a detailed chronology of the major breakthroughs and steps of the field, from the early beginnings in the late 1960s to the date of the article (2002). While it is hard to infer the taxonomic evolution until the third period of our analysis, part of their investigation confirms some of our most salient patterns: *"Late 1990s to early 2000s: Mutations are cloned and several genes that affect common processes are woven into molecular pathways"* — here, point (ii). Note that some other papers address and underline specific concerns of the third period, such as the development of comparative studies [4, 11].
3. Conference proceedings: Finally, some insight could be gained from analyzing the evolution of the session breakdown for the major conference of this community, "Zebrafish Development & Genetics" [8]. Topic distribution depends on the set of contributions, which reflects the current community interests; yet it may be uneasy for organizers to label sessions with a faithful and comprehensive name — *"organogenesis"* for instance covers many diverse subjects. Reviewing the proceedings roughly suggests that comparative and sequencing-related studies are an emerging novelty starting in 1998, at the beginning of the third period, which agrees with our analysis. On the contrary, the importance of issues related to the brain & the nervous system, as well as signaling, seem to be constant between the first and the third period, which diverges from our conclusions.

Note that the expert feedback here is obviously the most valuable, as it is the most exhaustive and the most detailed as regards the evolving taxonomy. The other sources of empirical validation are more subject to interpretation and a more comprehensive empirical protocol would have to include a larger set of experts, yielding more details as well as a more intersubjective viewpoint, thus objective.

Conclusion

In this paper, we proposed a method for creating a meaningful taxonomy of any knowledge community. After defining an epistemic community as the largest

group of agents using the same notions, we showed that GLs (concept lattices) automatically arrange a community into hierarchic fields and subfields, rendering overlaps among epistemic communities, commonly called interdisciplinary fields. Yet, since GLs organize the data but do not reduce it much (the set of all ECs can possibly be huge, thus intractable), we introduced criteria discriminating interesting ECs, therefore producing a partial epistemic hypergraph which is a *manageable* representation of the community hierarchical structure. A longitudinal study consequently made possible an historical description, by capturing stylized facts related to epistemic evolution such as field progress, decline and interaction (merging or splitting). We ultimately applied our method to the subcommunity of embryologists working on the model animal "zebrafish." Even with imperfect data quality (mostly due to weak linguistic assumptions), we successfully compared the results with expert-based taxonomies.

In other words, we designed a valid projection function from the low-level of relations between agents and notions, to the high-level of epistemological descriptions, thanks to GLs. So far, detection of this kind of community had been investigated both (i) in computer science [18, 23, 31] where the main drawback is the relevance for social science: clusters have an unclear connection with what social scientists would call communities; and (ii) in sociology, which by contrast introduce hypotheses and tools proper to social networks [5, 13, 37] and yield CMs more adequate to social group detection. Yet, most of these methods produce hierarchically structured clusters which are in fact more or less *dendrograms*: agents cannot be part of many non-embedded, overlapping communities, and are bound to belong to a lineage of increasing communities. This is easily solved by using lattices. Additionally, these categories would have been hard to detect using single-network-based methods, relying e.g. only on social relationships: agents of a same EC are not necessarily socially linked — single-mode data often implies massive information loss.

More generally, this kind of application of conceptual structures could be helpful to historians of science especially when there are massive amounts of data. The present study might be considered the first wholly non-subjective historical analysis of the "zebrafish" community. Also, GLs may be used in at least any semantic-community-finding case involving a relationship between agents and semantic items. As stated by Cohendet *et al.* [7], *"a representation of the organization as a community of communities, through a system of collective beliefs (...), makes it possible to understand how a global order (organization) emerges from diverging interests (individuals and communities)."* In addition to epistemology, scientometrics and sociology, other fields of application and validation include economics (companies and technologies), linguistics (words and contexts) and history in general (urban centers and industrial patterns [38]). Having significant results in many distinct fields would support the overall robustness of *GL-based compact taxonomy* building.

Acknowledgements. The author wishes to express the warmest thanks to Paul Bourgine and Nadine Peyriéras for the numerous and essential discussions and advices, as well

as three anonymous reviewers for their comments. This work has been partly funded by the CNRS.

References

1. R. Atkin. *Mathematical Structure in Human Affairs*. London: Heinemann Educational Books, 1974.
2. B. Berlin. *Ethnobiological classification - principles of categorization of plants and animals in traditional societies*. Princeton: Princeton University Press, 1992.
3. G. Birkhoff. *Lattice Theory*. Providence, RI: American Mathematical Society, 1948.
4. J. Bradbury. Small fish, big science. *PLoS Biology*, 2(5):568–572, 2004.
5. R. S. Burt. Cohesion versus structural equivalence as a basis for network subgroups. *Sociological Methods and Research*, 7:189–212, 1978.
6. M. Callon, J. Law, and A. Rip. *Mapping the dynamics of science and technology*. MacMillan Press, London, 1986.
7. P. Cohendet, A. Kirman, and J.-B. Zimmermann. Emergence, formation et dynamique des réseaux – modèles de la morphogenèse. *Revue d'Economie Industrielle*, 103(2-3):15–42, 2003.
8. Zebrafish development & genetics. Cold Spring Harbor, NY, 1994, 1996, 1998, 2000, 2001, 2002, 2003.
9. R. Cowan, P. A. David, and D. Foray. The explicit economics of knowledge codification and tacitness. *Industrial & Corporate Change*, 9(2):212–253, 2000.
10. H. Dicky, C. Dony, M. Huchard, and T. Libourel. ARES, Adding a class and REStructuring inheritance hierarchies. In *Actes de BDA'95 (Bases de Données Avancées), Nancy*, pages 25–42, 1995.
11. K. Dooley and L. I. Zon. Zebrafish: a model system for the study of human disease. *Current Opinion in Genetics & Development*, 10(3):252–256, 2000.
12. V. Duquenne, C. Chabert, A. Cherfouh, A.-L. Doyen, J.-M. Delabar, and D. Pickering. Structuration of phenotypes and genotypes through Galois lattices and implications. *Applied Artificial Intelligence*, 17(3):243–256, 2003.
13. L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40:35–41, 1977.
14. L. C. Freeman and D. R. White. Using Galois lattices to represent network data. *Sociological Methodology*, 23:127–146, 1993.
15. R. Godin, H. Mili, G. W. Mineau, R. Missaoui, A. Arfi, and T.-T. Chau. Design of class hierarchies based on concept (Galois) lattices. *Theory and Practice of Object Systems (TAPOS)*, 4(2):117–134, 1998.
16. D. J. Grunwald and J. S. Eisen. Headwaters of the zebrafish – emergence of a new model vertebrate. *Nature Rev. Genetics*, 3(9):717–724, 2002.
17. P. Haas. Introduction: epistemic communities and international policy coordination. *International Organization*, 46(1):1–35, winter 1992.
18. J. A. Hartigan. *Clustering Algorithms*. Wiley, New York, NY, 1975.
19. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.
20. A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323, 1999.
21. S. C. Johnson. Hierarchical clustering schemes. *Psychometrika*, 2:241–254, 1967.

22. J. T. Klein. *Interdisciplinarity: History, Theory, and Practice*. Wayne State University Press, Detroit, MI, 1990.
23. T. Kohonen. *Self-Organizing Maps*. Springer, Berlin, 3rd edition, 2000.
24. T. S. Kuhn. *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago, IL, 2nd edition, 1970.
25. S. O. Kuznetsov and S. A. Obiedkov. Comparing performance of algorithms for generating concept lattices. *Journal of Experimental and Theoretical Artificial Intelligence*, 14(2-3):189–216, 2002.
26. J. Lave and E. Wenger. *Situated learning: Legitimate peripheral participation*. Cambridge: Cambridge University Press, 1991.
27. L. Leydesdorff. In search of epistemic networks. *Social Studies of Science*, 21:75–110, 1991.
28. A. Lopez, S. Atran, J. D. Coley, D. L. Medin, and E. E. Smith. The tree of life: Universal and cultural features of folkbiological taxonomies and inductions. *Cognitive Psychology*, 32(3):251–295, 1997.
29. F. Lorrain and H. C. White. Structural equivalence of individuals in social networks. *Journal of Mathematical Sociology*, 1(49–80), 1971.
30. B. Monjardet. The presence of lattice theory in discrete problems of mathematical social sciences. Why. *Mathematical Social Sciences*, 46(2):103–144, 2003.
31. M. E. J. Newman. Detecting community structure in networks. *European Physical Journal B*, 38:321–330, 2004.
32. C. Roth and P. Bourgine. Epistemic communities: Description and hierarchic categorization. *Mathematical Population Studies*, 12(2):107–130, 2005.
33. G. Salton, A. Wong, and C. S. Yang. Vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.
34. R. R. Sokal and P. H. A. Sneath. *Principles of Numerical Taxonomy*. W.H. Freeman, San Francisco, CA, 1963.
35. G. Stumme, R. Taouil, Y. Bastide, N. Pasquier, and L. Lakhal. Computing iceberg concept lattices with TITANIC. *Data and Knowledge Engineering*, 42:189–222, 2002.
36. F. J. Van Der Merwe and D. G. Kourie. Compressed pseudo-lattices. *Journal of Experimental and Theoretical Artificial Intelligence*, 14(2-3):229–254, 2002.
37. S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, Cambridge, 1994.
38. D. R. White and P. Spufford. Medieval to modern: Civilizations as dynamic networks. Book Ms., 2006.
39. H. C. White, S. A. Boorman, and R. L. Breiger. Social-structure from multiple networks. I: Blockmodels of roles and positions. *American Journal of Sociology*, 81:730–780, 1976.
40. R. Wille. Restructuring lattice theory: an approach based on hierarchies of concepts. In I. Rival, editor, *Ordered Sets*, pages 445–470. Dordrecht-Boston: Reidel, 1982.
41. R. Wille. Concept lattices and conceptual knowledge systems. *Computers Mathematics and Applications*, 23:493, 1992.