NEST STREP proposal for the Pathfinder program “Tackling Complexity in Science” *(first draft)*

CAS²: a Complex Adaptive System for a Complex, Adaptive Science

Some other possible titles

BOOT-CS; bootstrapping complexity science
COMACOSC: a complex map of a complex science

Consortium Members

- VUB: Vrije Universiteit Brussel, Center for Evolution, Complexity and Cognition, Belgium (director: Francis Heylighen) (Coordinator)
- ISI: Institute for Scientific Interchange Foundation, Multiagent Division, Italy (director: Sorin Solomon)
- CUE: Crakow University of Economics, Poland, ... research group (director: Czeslaw Mesjasz)
- TIFR: Tata Institute of Fundamental Research, Homi Bhabha Center for Science Education, India, ... team (director: Nagarjuna G.)
- LANL: Los Alamos National Laboratory, Knowledge Systems and Modelling Research Team, USA (director: Cliff Joslyn)

Background and motivation

As production, transport and communication become ever more efficient, we interact with ever more people, organizations, systems and objects. And as this network of interactions grows and spreads around the globe, the different economic, social, technological and ecological systems that we are part of become ever more interdependent. The result is an ever more complex "system of systems" where a change in any component may affect virtually any other component, and that in a mostly unpredictable manner.

The traditional scientific methods, which are based on analysis, isolation, and the gathering of *complete* information about a phenomenon, are not ready to deal with such complex interdependencies characterized by non-linear interaction and unpredictable change. The emerging science of complexity offers the promise of an alternative methodology that would be able tackle such problems. One of its core concepts is that of a *complex, adaptive system* (CAS). This is a system consisting of a variety of agents (which can be complex and autonomous systems themselves), connected by a tissue of interactions. This network of interactions typically *self-organizes*, adapting its structure to the changing environmental constraints, while trying to maximize the benefits for the individual agents. Examples of complex adaptive systems are markets, ecosystems, multicellular organisms, brains, and dissipative chemical reactions. CAS are typically studied by means of a multi-agent computer simulation, where software agents following simple rules are allowed to interact with each other and their environment. These interactions typically give rise to emergent patterns of activity that are difficult or impossible to predict from the properties of agents individually.

Developments such as these have created the hope for a multitude of applications of the complexity paradigm, in a variety of domains characterized by complex interactions: ecology, management, information technology, molecular biology, medicine, politics and governance, culture, and sociology. However, complexity science itself is very complex and in full evolution. That makes it very difficult to develop good applications.

The scientists using complexity concepts typically fall into two groups:

1) the *specialists*, who are expert in specific mathematical or computational modelling methods—such as chaos theory, networks, or multiagent simulations. They further develop their expertise by applying these methods to carefully chosen simulations or data, typically from artificially restricted domains. However, they generally lack experience with the unrestricted complexity of real-world systems which is difficult to capture in a specific mathematical formalism.

2) the *practitioners*, who are confronted with very complex situations extending over multiple domains, such as the interplay between economic, social, technological and biological factors in health-care management. They are motivated to try out new complexity ideas, but typically lack the expertise to choose and apply the most appropriate concepts, models or methods for the problem at hand. Therefore their analyses rarely go beyond the level of metaphor and analogy.

What is lacking is an integrated body of knowledge, a comprehensive, coherent and transparent system of concepts, principles and methods. Given the great diversity of contributions to complexity science coming from very disparate disciplines, and the constant development of new ideas, no single person or even organization is able to provide such an overview of complexity knowledge. This problem has been tackled to some extent by the creation of networks of researchers and institutions—such as the EXYSTENCE and ONCE-CS networks funded by the IST-FET program, and the GIACS network funded by NEST—that promote collaboration and information exchange. However, with dozens of participating institutions, working in different countries, disciplines and traditions, it is very difficult to coordinate the activities in such a way that they would produce a coherent, integrated body of knowledge.

To achieve this, we need more than the traditional "social networking" methods of conferences, workshops, education programmes, student exchanges, etc. We need tools for the collective creation, representation and management of complex knowledge. While software systems for knowledge management and collaboration have been developed in the last few years, these tools typically start from relatively simple assumptions about what a knowledge system is and how it can be developed. Especially with knowledge as complex, ill-structured, and rapidly evolving as the one in complexity science, we need a more adaptive approach.

This new paradigm for knowledge management may be found in complexity science itself. The core idea is to design our integrated knowledge system as a CAS: a network of interacting knowledge components that self-organizes while constantly adapting to the changes in its environment. That is why we have called our proposal CAS^2, i.e. applying CAS knowledge to itself, thus "bootstrapping" the development of the complexity field. This should lead to an "intelligent" self-organizing system for knowledge representation, distribution and management, that would capture all the major concepts, methods, resources and applications of complexity in a format that is easy to use by everybody.

This system could then provide the "missing link" between the other complexity related projects, so that any researcher working on any aspect of complexity would immediately be able to find out how her approach fits in with other approaches.

Objectives

The present project proposes to create a complex, adaptive system of knowledge about complexity and its potential applications to science and society.

This system will be structured as a *hybrid network*, consisting of different types of nodes and links. Nodes are realized as hypermedia documents that contain easy-to-understand and accurate information about a particular topic. Topics can be concepts, methods, models, publications, authors, institutions, applications, software systems, etc.; in other words, any type of resource that can be of use for a scientist wishing to tackle a complexity-related problem. A node about a concept (e.g. chaos, self-organization, power law, percolation, ...) would typically contain a short definition with some basic illustrations. A node about an author may contain a short biography, affiliation, domain of expertise, and contact information. Nodes are connected by hyperlinks of different types. For example an author may be expert in a particular model, creator of a software application and member of an institution, and the node representing that person would therefore point via appropriately labeled links to all these associated nodes. Similarly, a concept may be a special case of another concept, applicable to a particular domain, and described in one or more publications, linking to the corresponding nodes.

This network is intended to have the following characteristics:

• **distributed**: the data of the network will be spread over different computers communicating via the Internet. They will be accessible at any moment from any place using a standard web browser. Different contributors in different places can add or edit the content of the nodes independently. There is no central control: no single individual, institution or computer is in charge of the overall network structure (although a board of supervisors will monitor its general quality, having the possibility to intervene if additions are really substandard). The global structure of the network is expected to emerge from the local (inter)actions of many different contributors.

• **multilevel:** the network will offer different levels of granularity or abstraction, so that people who wish to get a general view of a particular domain don't need to get into the details, while still being able to zoom in on particular aspects that have drawn their attention, and then zoom out again to see how a specific detail fits into the larger picture.

• **context-dependent perspectives:** different users will get different (partial) views of the same network depending on their present interests or focus. Since the network will be far too complex to show as a whole, or even as a local section, the system needs to be selective about which nodes and links it presents to each user, keeping aspects that are less relevant in the background. By thus adapting the perspective to what is most important for the user and the situation, the system should avoid the problems of cluttering, confusion, or the feeling of being “lost in hyperspace” that is typical for large hypertext networks such as the Web. But any change in the focus should lead to an immediate shift in the view, so that aspects relevant to the new focus are revealed.

• **self-organizing, adaptive, learning**: not only the perspective offered on the network, but the network itself should constantly adapt to the users' actions. Registered users can individually add, delete or edit nodes and links (although a version control system will always allow to reinstate earlier versions). The system will moreover adapt to the collective activities of the users, e.g. by strengthening frequently used links or clustering nodes that tend to be used together. By monitoring which nodes tend to be consulted in which order, the system will become better at anticipating the users' desires, offering shortcuts and pointing them to the nodes that are most likely to be relevant for their personal interests. At a higher level, the system will create emergent structures, by analysing which regularities are implicit in the pattern of linking or browsing, representing these by new nodes at a higher level. The system will also try to determine the gaps in its own knowledge, i.e. the nodes or linking patterns that are incoherent or incomplete, and point these out to the users most likely to be expert about this missing knowledge. As such the knowledge system will continuously improve itself, not only by capturing the knowledge implicitly held by its users, but by finding larger, more abstract patterns and connections between fragments of knowledge, and thus possibly generating novel insights.

• **mediating between sources and users of knowledge:** the network will not only enhance its own structure, but the structure of the implicit social network formed by its users, thus bridging the gaps between groups, individuals, or communities who don’t know about each other, yet are interested in related topics. For example, the system may note that there is a substantial overlap between the usage pattern of a mathematician looking for social applications and a social scientist looking for mathematical methods, and point the users towards each other. Thus, the system should function as a mediator, bringing users in contact with other users with similar or complementary interests, thus promoting synergetic interaction and collaboration between the members of the community.

Individual workpackages and deliverables

Although the way we have presented our overall objectives may seem very ambitious, and critics may wonder whether such a level of "intelligent" self-organization is practically achievable within the limited duration of the project, we believe that such a system can be built through a number of relatively independent workpackages (WPs). Even if some of these workpackages would fail to deliver on their promises, the integration of the results from the other WPs should still be sufficient to create a system much more powerful than anything that exists at the moment. Another reason for us to be ambitious is that the different consortium partners have already performed quite some preliminary research, and gathered a lot of expertise in each WP domain. The novelty of the present project is the integration of all the WPs, thus producing a system that is much more than the sum of its parts.

WP 1: a semantically structured database for collaborative knowledge management

Partners: TIFR, with input from VUB, ISI and LANL.

This package will further develop and extend the GNOWSYS system of which a first prototype has been implemented at TIFR. This knowledge management system is based on a semantic network formalism, where nodes and links belong to a limited set of predefined types with specific formal properties (e.g. transitive, antisymmetric or one-to-many). Although the system uses its own distributed database engine, its nodes can be consulted and edited via the web, providing both HTML and XML formats. The software is freely available and open source, under the GNU licence. This means that any present or future contributor to the complexity network will be able to install and if necessary customize the software, without cost or restriction. To fully support the CAS^2 project, the system will need to be extended with wiki and other groupware functions, so that its content can be easily edited remotely by non-expert users. Moreover, its interface will need be streamlined so as to be as user-friendly as possible (see WP 10).

WP 2: an ontology for the complexity domain

Partners: VUB, LANL, TIFR

The present GNOWSYS prototype incorporates a basic ontology of node and link types, which encompasses such fundamental categories as objects, processes, and relations.

However, to create a detailed and coherent map of the whole complexity domain, we will need to develop a much richer ontology of classes, subclasses and relations between classes, that can unambiguously capture all the basic types of complexity knowledge. Given that the complex systems approach extends over practically all traditional disciplines, this ontology should be *transdisciplinary*, being equally applicable to many different applications and types of systems. This will require the compilation of a list of basic categories, the resolution of ambiguities (e.g. when a term in one domain has a different meaning in another domain, or when two different terms actually refer to the same type of phenomena), and the determination of the precise dependencies between the categories.

For example, our ontology will distinguish at the most abstract level between types (abstract classes) and tokens (concrete instances). The token category will distinguish between objects, events and processes. Objects may be subdivided in, among others, persons, organizations, places, and physical objects, whereas organizations may subdivide further in universities, research institutes, associations, etc. For these token classes our ontology may build further on existing ontologies that have already been developed in other domains. But for the abstract types, the complexity perspective will require the disambiguation, classification, and organization of much more complex and less intuitive concepts, such as self-organization, dynamical system, order parameter, agent or autocatalysis. For example, the ontology may classify autocatalysis as a_part_of self-organization, but also as a_kind_of feedback, which itself is a_kind_of process.

A coherent classification of these abstract, and often not yet very well understood concepts will on its own already be an important step towards the integration and explicitation of complexity knowledge. This ontology will be implemented in the GNOWSYS system as a labelled, directed network of "empty" nodes that designate the basic categories, forming a skeleton for the knowledge network being built.

WP 3: a knowledge map of complexity concepts

Partners: VUB, CUE, ISI, with input from TIFR, LANL

Given an empty skeleton of linked ontological categories (WP 2), the next task is to "fill" the nodes with content, that describes the node in a more intuitive and concrete way. This content will typically consist of free format text, supplemented where necessary by pictures, diagrams, formulas, code or even demonstrations. At the least, a node should contain a basic definition of the concept, which is as clear and understandable as possible. Any word in the definition or explanation that may not be obvious will be hyperlinked to the node explaining that concept.

Initially, we will use two basic methods to provide content. The simplest is to gather existing, freely available information. For this we have a multitude of sources at our disposal that are either public domain (e.g. the extensive Wikipedia and other dictionaries, glossaries or encyclopedias available over the web) or to which the consortium members have privileged access because of their earlier activities in this domain (e.g. the Principia Cybernetica Web with its Web Dictionary of Cybernetics and Systems, the database collected by Complexity Digest, and the materials produced in the framework of the GIACS coordinating action). However, given that the format, style and quality of these sources can be disparate, this will require extensive manual editing to fit the material in a coherent framework. This editing task will be distributed over the different individual collaborators in the consortium.

The second method to produce content is simply to write the texts ourselves, using our own extensive knowledge and experience in the domain. This will be necessary in those cases where pre-existing content is either lacking or of insufficient quality.

A third method is relegated to a later stage (WP10), when outside researchers are invited to contribute their expertise to the network. However, to ensure that this will happen smoothly, we will need to maximally prepare the terrain in the present stage, by providing plenty of examples filled with high quality content complementend by largely empty "stubs", i.e. minimal characterizations of important ideas, already linked up to the most relevant ideas, so as entice others to fill in the lacking details.

Although the creation of content (WP3) logically appears to follow the definition of an ontology (WP2) the two workpackages will be started in parallel, since the gathering of existing definitions may help us become aware of possible ambiguities and ways to resolve them in the ontology. In any case, the system should be flexible enough to allow the introduction of as yet unstructured content nodes, where their position in the ontology is defined later on, if ever. The moral is that it is better to have an intuitive, albeit ambiguous characterization of a concept than no characterization at all. Moreover, the methods of network self-organization described in WPs 8 and 9 may help us to disambiguate equivocal or vague concepts while suggesting a clear location for them in the ontology.

WP 4: a knowledge map of concrete complexity resources

Partners: CUE, ISI, LANL, VUB with input from TIFR

This workpackage will perform the same function of collecting content as WP3, but now for concrete resources rather than abstract concepts. This means that we basically want to develop a detailed map of the complexity community, its tools, products and activities. These data will be fitted into the predefined ontological categories, such as person, institution, conference, paper, book, etc. As noted earlier, these "token" nodes will also be linked into the "type" network produced by WP3, so that a node representing a paper on the subject of non-linearity will be linked to the node defining the concept non-linearity, and vice-versa.

To collect these data we will also use existing sources, such as the data gathered by other European coordination projects (GIACS, ONCE-CS...), existing associations and networks, such as the one organized by the Center for Complexity Research in Liverpool University, bibliographies and lists of journals and societies collected by the Principia Cybernetica Project or Complexity Digest, etc.

To work efficiently, in addition to manual entering and editing, this WP can use automated programs ("webcrawlers") that mine the web for specific types of data, and convert these to a common format. For example, a web crawler may download all the abstracts of papers and preprints available on the web that include one or more complexity-related keywords (self-organization, complex systems, ...), compile the references into bibliographies, the authors into lists of complexity researchers, and their institutions into complexity centers. This program can use simple heuristics, such as including an author into the database if it finds at least three complexity-related papers by that author, or including an institution if it finds at least two complexity authors affiliated with that institution... Similarly, a paper could be deemed complexity-related if it contains a minimum number of complexity-related keywords, or if it is referenced by at least three other complexity-related papers.

Such a program does not need to make a sharp "yes-no" decision as to whether a particular paper, author, or conference belongs to the complexity field, but simply give them a "complexity-relevance score" that takes into account a multitude of heuristic factors such as the ones mentioned. Since the database will have enough space to include thousands or even millions of nodes, there is no need to make a strict selection. However, to use such an ever larger and more complex web of data efficiently, we will need good methods to find the most relevant nodes, taking into account both their general relevancy score and their association with the more specific interests of the user. This forms the subject of WP6.

WP 5: a knowledge map of complexity objectives and application domains

Partners: CUE, ISI, VUB, with input from TIFR, LANL

A third, smaller subnetwork can be very useful to the strategic development of complexity science. It would list not so much (abstract or concrete) resources, but objectives or applications for which these resources could be used. Its nodes would describe various unsolved problems and potential application domains that might be tackled with the help of complexity science, with direct links to the concepts, methods, models or resources most likely to be relevant. The network can be structured quasi-hierarchically, with the higher levels consisting of very broad, abstract applications domains, such as "ecological problems" or "societal problems", and in the lower levels specific cases such as "modelling breast tumour as a complex system" or "predicting failures in electricity distribution networks". Moreover, nodes from this network may be labeled with a perceived level of importance, urgency or difficulty, so that it becomes possible to focus e.g. on the problems that are most important and urgent to tackle, but that demand least effort.

This network can again be compiled from existing sources, such as position papers and draft "roadmaps" produced by the complexity community and supplemented by the members of the consortium from their own experience or from the suggestions of invited experts.

WP 6: a context-dependent search/recommendation algorithm for complex networks

Partners: VUB, LANL with input from ISI

Whereas the previous WPs basically apply traditional methods of knowledge acquisition and representation in a systematic manner to the complexity domain, the next WPs apply ideas from the complexity paradigm itself to manage and organize the resulting complex network.

WP 6 intends to further develop and apply a promising network exploration method that was recently discovered by researchers at VUB and LANL: *particle diffusion*. Traditional algorithms to find information in networks, such as depth-first search or spreading activation, use continuous or deterministic methods to systematically investigate *all* nodes. In a network where nodes of different degrees of importance are constantly added, changed or repositioned by independent parties, such an exhaustive search is likely to be inefficient and unreliable. In contrast, our approach—like most CAS models—is based on stochastic and discrete methods, that let the solution "self-organize".

Search is performed by extremely simple agents, which we call "particles". Particles move from node to node along the links, while keeping track of where they have been. Particles choose a link at random from the ones available at the present node, with a probability proportional to the strength (importance) of the link. They can also be programmed only to follow links of a certain type (e.g. "refers_to", but not "is_authored_by"). The initial nodes from where the search starts are "seeded" with a large number of particles, becoming sources for particle diffusion. Potential solution nodes (e.g. papers with the keyword "scale-free") act as sinks, keeping the particles they receive, without letting them diffuse further. At the end of the process, the nodes that have gathered most particles are offered as solutions to the query, ranked according to their particle counts.

The initially seeded nodes define the *context* of the query, i.e. the present focus of interest of the user who initiates the query. This can be determined automatically and implicitly by registering the last nodes visited by that user, allocating them a number of particles proportional to the interest expressed by the user (e.g. the time spent reading the node), and the recency of the visit. Without further specification of a query, particle diffusion from these context nodes will bring forward those nodes most related to the whole of nodes that were a focus of interest. Thus, the user will constantly receive tailor-made recommendations of those nodes that seem most relevant for his/her present interest-profile, without needing to explicitly formulate queries.

However, if the user wants to search for something with specific characteristics (e.g. papers by authors working in Poland on the subject of scale-free networks), the diffusion process can be constrained to only allow particles to remain in nodes that fulfil these requirements. Even then, the ranking of the retrieved nodes will continue to reflect the implicit context. For example, if the user on previous visits showed particular interest in social networks, papers on these types of networks will be recommended more highly than those on e.g. genetic networks.

This WP will experiment with different variations on the basic particle diffusion algorithm, and try them out in different subnetworks and the complexity network as a whole, for different types queries, in order to find the most efficient approach.

WP 7: a visual browser for complex networks

Partners: ISI, TIFR with input from LANL, VUB

Organizing query results as a relevancy-ranked list, as proposed in WP6 or by traditional search engines on the web, is probably the simplest and most intuitive way to represent complex data in a user-friendly format. However, such a linear presentation does not do justice to the complexity of the underlying network of data and associations. Many attempts have therefore been made to develop *graphical* representations of networks that show the many paths along which one node can connect to the others in its neighborhood. However, the fact that such network visualizations have never become popular as a way to navigate the web points to an intrinsic shortcoming: as soon as the network contains more than a few dozen nodes and links, the visual representation starts to look like "spaghetti", proposing a clutter of cross-cutting lines that confuses more than it illuminates.

Several complexity-related algorithms, such as relaxation, have been proposed to reduce the clutter by positioning nodes and links with as little overlap as possible while trying to make the distances reflect the degree of relatedness. The ISI group has used such methods to develop the Postext system, which provides a visual interface for the browsing and editing of complex hypertext networks. However, to better control clutter we need to integrate these methods with the methods used in WP6 to determine contexts.

The principle is that any user at any moment is characterized by a unique context or focus of interest, determined by his or her previous actions. As explained in WP6, visited nodes are seeded with particles, which then diffuse to activate neighbouring nodes, resulting in degrees of activation that reflect the user's assumed degree of interest in these nodes. The number of particles travelling along a certain link similarly reflects the relative importance of that link in the present context. This allows us now to create a context-dependent snapshot of the network, showing only the few dozen nodes and links through which most particles have passed, with the salience of the component proportional to the number of particles. When the user selects one of these nodes for further inspection, the context shifts, and extra particles are made to diffuse from the chosen node. As a result, the view too shifts, and nodes and links that were pushed into the background may now gather enough particles to become visible, while other nodes may now lose enough particles to disappear from view.

The result of this WP will be the development of a visual browser that at any moment shows the neighborhood of nodes and links that is most relevant for the user's present interest. The view thus constantly adapts to the user's needs, taking in any section of the network that seems relevant, without ever becoming too complex for quick visual inspection. The browser will also help the user to edit the network: within the graphical representation, nodes and links can be created, deleted or opened up for textual editing by simple clicking.

WP 8: algorithms for link learning in a complex network

Partners: VUB, LANL with input from ISI

This WP proposes a first step in the *self-organization* of the knowledge network: the unsupervised creation by the system of new links between nodes, and the adaptation of their weights to the way they are used. The rationale is that in a system with many thousands of nodes, and millions of potential links, no one can have enough knowledge to determine the optimal linking pattern: this pattern will have to emerge out of the collective activities of all users.

A simple method to achieve this was discovered several years ago at the VUB, and tested out in collaboration with LANL. The inspiration comes from two self-organizing systems studied by complexity science: the brain, and ant colonies. In the brain, the connection between two neurons (the synapse) is strengthened whenever these two neurons are activated within a short interval (Hebbian learning based on co-activation), while weakening otherwise. When ants discover a source of food, they leave a trail of pheromones on the path from the food back to the nest. The pheromones increase the probability that other ants would follow the same path and thus strengthen the trail with more pheromones, leading to a quick non-linear growth (autocatalysis). When paths are no longer used, pheromone concentrations quickly decay. Moreover, shorter, more efficient paths will gather more pheromones, and thus eventually win the competition with longer paths. In both cases, neurons and pheromone trails, the result is a self-organizing, adaptive network where successful connections are strongly amplified, while unsuccessful ones are erased.

The application to a knowledge network is straightforward: whenever a user shows interest for a node A, and shortly afterwards for a node B, the link between A and B is reinforced—or a new link is created if none existed yet. Unused links, on the other hand, slowly decay and eventually disappear. This further allows the system to discover *shortcuts*: if A is linked to B, and B to C (D, E, ...), then the system may directly link A to C (D, E, ...) if its turns out that people who like A generally also like these subsequent nodes. The result is that eventually every node A will be linked to all (and only) those other nodes that are most likely to be relevant for someone interested in A. These links are presented in the order of relevance, i.e. the ones that gathered the strongest amount of "pheromones" or "co-activation" first. This optimises users' traversal of the network, minimizing the number of detours or trial-and-error explorations that they need to do before finding the most interesting nodes.

Moreover, the resulting pattern of weighted links directly facilitates context-dependent recommendation (WP5) and visualisation (WP6). A cloud of particles, representing not just the present node but the previous context of browsing, will diffuse according to the links and their respective weights. The better these links represent the actual associations between nodes, as experienced by users, the better the recommendations found in this way. Thus, we may expect that as the network learns a more optimal linking pattern, recommendations and visualizations of context-relevant nodes will become increasingly accurate.

WP 9: algorithms to support hierarchical self-organization of a complex network

Partners: VUB, LANL with input from ISI

The next step in making the network more adaptive, is to enable the self-organization of not only links, but nodes. The idea is that if several nodes are closely related, as represented by strong and dense links between them, then these nodes can be seen as a single *cluster*, to be represented by a higher order node. For example, if the system would discover that a number of researchers regularly collaborate or cite each other's papers, while these papers link to several of the same concepts, then it may recognize them as a *research community* with the combination of these concepts as research subject. It can then create a "stub" node representing this community, linking it to all its members and its most representative papers and concepts. Moreover, the system may invite these authors or other interested parties to read, correct and complete the information in this node, so that later visitors will get an accurate description of this new community. The advantage is that users interested in this emerging new research theme will immediately get a list of all the most important people, publications, etc. that belong to it even before the community has become formally recognized, and won't have find them by browsing, following links from the one to the other, without getting a clear picture of how they fit together.

The process of creating a new node representing a cluster of associated nodes can be applied *recursively*: cluster nodes may themselves be clustered with other nodes, forming subsequent levels of "clusters of clusters". This may result in the automatic generation of taxonomies or classifications of themes, communities, application domains, etc. This "self-organized" hierarchical organization may extend and improve the hierarchical structure created by the authors of the network, thus supplementing their limited knowledge of the complex relationships that exist between nodes in the network.

There are many algorithms to perform cluster analysis. However, most of those (such as the traditional k-means) make quite restrictive assumptions about the size or shape of clusters, and the nature of the connections between their members. To efficiently harvest the emergent order in a fully self-organizing, complex network, we should ideally develop a new method. Very recently, researchers from VUB and LANL have started to explore a new algorithm, building on our methods of particle diffusion (WP6) and Hebbian learning of links (WP7).

The main innovation is that a new node is generated whenever a collection of nodes are co-activated (e.g. consulted by the same user, or selected by the same cloud of diffusing particles), and when this cluster of activation cannot be accounted for by existing, hierarchically superior nodes. Since many of the nodes in this temporary cluster may be co-activated purely by coincidence, the cluster needs to be further refined. This can be achieved through Hebbian learning. Whenever there is a substantial overlap between the initial cluster and a later pattern of activation, the links between individual nodes and the cluster node become differentially reinforced: the links that are repeatedly used become stronger, those that are used only sporadically decay and die off. This will only leave those nodes in the cluster that are frequently co-activated in different contexts.

WP 10: testing and optimising the system

Partners: ISI, TIFR with input from VUB, LANL.

Assuming that the previous workpackages have been successful, we will have developed a complex knowledge management system including various advanced tools for navigation, visualisation, recommendation, and self-organization. This system is intended to make it easier for its users to find the information they need and to improve their understanding of how complexity concepts can help solve problems. However, the danger is real that this system will merely add another layer of complexity to the already complicated and confusing landscape of complexity science and its associated networks. Ironically, many information systems created to facilitate interaction merely confuse their users, so that after a few failed attempts they simply stop interacting with it, sticking to the less powerful systems that they know how to use. To avoid this danger, we will design an interface that is as user-friendly and intuitive as possible. To achieve that, we can rely on different principles.

The first principle is to *keep it simple*, i.e. minimize the number of icons, commands, menus, decorative elements, logos, etc. that the user needs to assimilate before being able to make a choice. The second principle is to design an effective *default hierarchy*: the default option—the one that applies without specific decision by the user—should always be the one that most people are likely to use; more advanced options can be chosen via an additional step, such as a pop-menu, or the selection of a link announcing more choices. For example, in the Google search engine, the default action is to search the web, and additional possibilities are announced via a minimum of links pointing to web pages with more advanced functions, so that the default start page is minimally cluttered. The third principle is to make use of the expertise of researchers specialized in *human-computer interaction,* who have spent years observing and experimenting with people using different types of visual, spatial and verbal configurations to see which ones work best. To achieve that, this WP will invite a number of selected HCI experts to advise us in the creation of our interface, and to afterwards examine and criticise the system.

The fourth principle, finally, is to test out the system with actual people, using objective, quantifiable criteria to measure its performance. A key criterion is how satisfied users are with the recommendations made by the system. To measure this, during the test phase we can simply insert a small 5-point scale e.g. at the bottom of a page, where users can mark the option that best describes their degree of satisfaction with the present information, e.g. from "useless" to "just what I needed". This will allow us to check whether the different novel features, such as "cluster nodes", really offer any improvement over the manually created hyperlinks. Moreover, this feedback signal allows us to optimise the system by tuning the various parameters (e.g. the number of "particles" given to various contextual nodes) so as to minimize the difference between the actual degree of satisfaction and the degree of relevancy as estimated by the system.

WP 11: creating a community of users and contributors

Partners: VUB, CUE, ISI, with input from TIFR, LANL

For the final workpackage, we assume that we have a well-designed and tested system packed with a variety of relevant information and augmented with different tools and algorithms to cope with the intrinsic complexity of the knowledge domain. The next step is to involve the people for whom it is intended, i.e. the different communities of scientists and practitioners interested in understanding complexity methods. These people should not only know that the system exists, or merely know how to use it, but actively contribute to its content, thus making it expand, learn and develop. A model to keep in mind here is the Wikipedia community, which in a few years has made the Wikipedia knowledge base into the largest encyclopaedia in existence.

To achieve that, we will start with targeted and repeated publicity actions towards the communities most likely to profit from the system. For example, we will send email announcements to various complexity-related mailing lists, create links to the system in complexity-related websites, present talks and seminars about the system at complexity-related meetings and conferences, use our many contacts around the world to personally invite people to contribute, publish papers and reports describing the system, etc. Given our experience with similar enterprises, such as Principia Cybernetica and Complexity Digest, we expect this process to quickly become self-amplifying: the more people know about the system, the more references will be made to it in various publications, websites, discussions, etc., and thus the more people will get to know about it.

The most reliable indicator of success in this respect is a website's PageRank, the measure used by the Google search engine to determine how authoritative a page is. If you search for a particular keyword, e.g. "self-organization", in Google, the first pages you get will typically belong to websites with a high PageRank, such as Wikipedia or Principia Cybernetica. The quickest way to increase a website's PageRank is to have it linked to by such websites that already have a high PageRank. This is easy to achieve for us, given that we have direct access to a number of these websites, such as Principia Cybernetica or Complexity Digest. But the main criterion to achieve a high PageRank and thus a high degree of visibility and user participation is, of course, *quality*: the better (i.e. more reliable, informative, easy to understand, novel, applicable, ...) the information offered, the more people will use it, and the more attractive it will become to further people.

In that way we hope to create a system that reaches an ever expanding audience of people across Europe and the world who are interested in understanding and applying complexity science. That in turn will boost the development of complexity science, as a constant input of fresh ideas and motivated contributors, together with a system that helps organize these contributions into a clear and coherent framework, will produce ever more broad and convincing models and understanding of the complex world in which we live.