

Classification, thésaurus, ontologies, folksonomies : comparaisons du point de vue de la recherche ouverte d'information (ROI).

Manuel Zacklad

Université de Technologie de Troyes

Equipe **Tech-CICO** (Technologie de la Coopération pour
l'Innovation et le Changement Organisationnel)

UTT ICD/Tech-CICO - FRE CNRS 2848

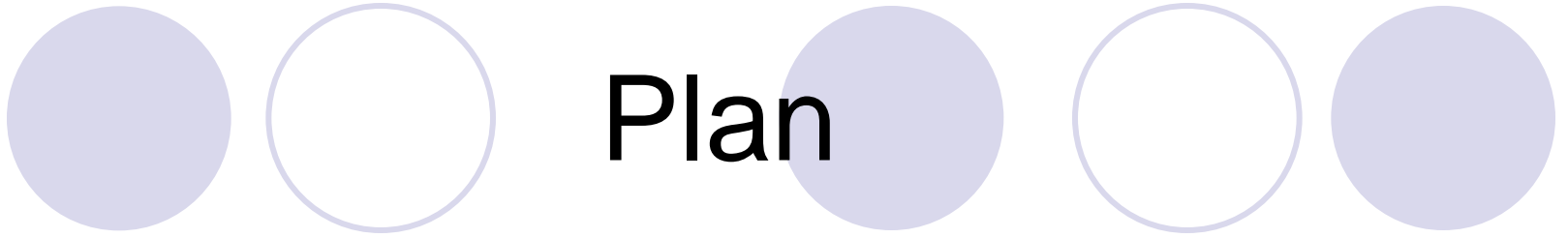


Objectifs

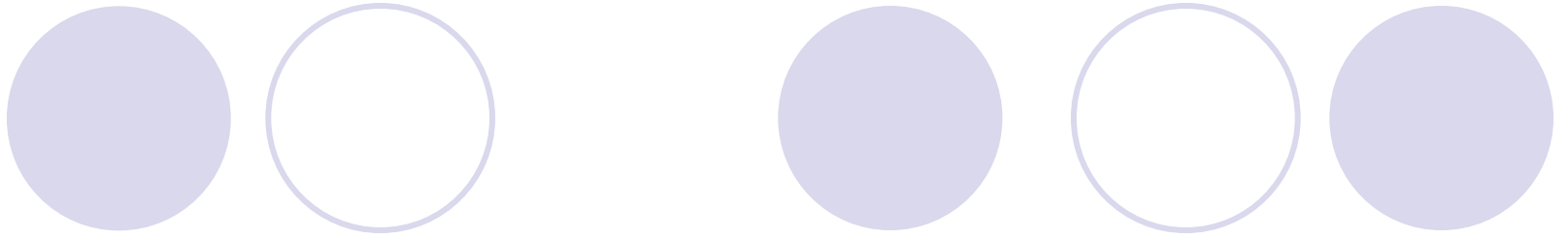
- Elaborer des critères d'évaluation des différents Systèmes d'Organisation des Connaissances (SOC/KOS): classification, thésaurus, ontologie, folksonomie
- Sur la base de ces critères, positionner le langage HyperTopic que nous élaborons à Tech-CICO (Université de Technologie de Troyes) pour développer notre vision du Web Socio Sémantique et les méthodes de « Recherche Ouverte d'Information »
- Concevoir des méthodes et des plateformes informatiques tirant parti de la complémentarité entre les différents SOC

Caractériser les fonctionnalités des SOC

- Les SOC ont un rôle essentiel pour identifier les informations documentaires pertinentes via la navigation (browsing)
- Mais le cadre conceptuel « informatique » de la Recherche d'Information (information retrieval) est insuffisant pour qualifier pleinement les fonctions des SOC
- Proposition d'un cadre conceptuel plus large, la *Recherche Ouverte d'Information* (ROI), lui-même inséré dans une description plus générale des activités d'enquêtes (résolution de problème)



- La recherche ouverte d'information dans le contexte des démarches d'enquête
- Présentation des différents systèmes d'organisation des connaissances
- Comparaison du point de vue de la ROI



Recherche ouverte d'information et démarche d'enquête

Recherche d'information vs Recherche ouverte d'information (information retrieval vs open information research)

- RI classique : **accès** à un document numérisé dont la localisation n'est pas connue mais dont l'existence ou la pertinence est acquise.
- Recherche Ouverte d'Information liée à l'analyse d'une situation complexe sans savoir à l'avance s'il existe des ressources documentaires susceptibles de répondre au besoin : *Recherche Ouverte d'Information*
- Processus de découverte et d'apprentissage permettant de poser un problème dans le cadre d'une démarche d'enquête cf. les phénomènes de sérendipité (découverte fortuite) associés à la RI,
 - « serendipitous information retrieval » Toms (2000)
 - « processus d'apprentissage dynamique » associés à la RI (Ertzcheid et Gazzelot 2003) qu'ils souhaitent, comme nous, plus ouverte aux phénomènes de complexité et de créativité.

RIO et démarche d'enquête

- Enquête (au sens de J. Dewey) deux phases principales:
 - La première est une phase de *recherche d'information*, ou de compréhension, permettant l'analyse de la situation, des options, des ressources disponibles.
 - La seconde est une phase de *mise en œuvre*, de traitement, de décision, de synthèse.
- Pas d'ordre strict, correspond également à deux dimensions de l'enquête qui s'entremêlent de manière étroite
 - Les phases de mise en œuvre sont également une manière d'acquérir de l'information qui contribue à la RI.
- Deux types de RI
 - **intra-documentaire** (situation déjà documentée)
 - **extra-documentaire** (observation directe de la situation et de ses ingrédients ou interaction avec des personnes susceptibles d'apporter des réponses aux interrogations rencontrées)

Démarche d'enquête

**Phase de
Recherche
Ouvverte
d'Information**
(on parlera de RI pour
faire court)

**Phase de mise en
œuvre**

Caractère extra-documentaire ou intra-documentaire de l'enquête

Première typologie des enquêtes	Mise en œuvre exprimée de manière non documentaire	Mise en œuvre exprimée de manière documentaire
Recherche d'information extra-documentaire	<i>Extra-documentaire pur</i> Les sources primaires sont extra-documentaires et la mise en œuvre est extra-documentaire : p.e. choisir un ami pour l'inviter à une sortie	<i>Extra /Intra-documentaire</i> Les sources primaires sont extra-documentaires mais la mise en œuvre est de nature documentaire : p.e. rédiger le compte rendu de plusieurs visites d'appartements
Recherche d'information intra-documentaire	<i>Intra /Extra-documentaire</i> Les sources primaires sont documentaires mais la mise en œuvre est extra-documentaire : p.e. faire un voyage lointain après s'être documenté	<i>Intra-documentaire pur</i> Les sources primaires sont documentaires comme la finalité qui se traduit par une mise en œuvre documentaire : p.e. rédaction d'un article à partir d'autres articles

Typologie des enquêtes inspirée de la psychologie cognitive

- On y distingue trois types d'enquêtes selon leur finalité principale :
 - Enquête de localisation et d'accès : la situation problématique est générée par l'absence d'un objet connu
 - -> trouver la localisation de cet objet, y accéder ou l'acquérir.
 - Enquête visant une sélection ou une décision : absence d'un objet dont on connaît l'existence parmi d'autres mais qui n'a pas encore été identifié
 - -> sélectionner l'objet souhaité au sein d'un ensemble et y accéder ou l'acquérir.
 - Enquête visant une synthèse ou une conception : absence d'un objet n'ayant pas encore d'existence
 - -> synthétiser ou de concevoir cet objet.

Phase de l'enquête / Type d'enquête	RI documentaire	RI extra-documentaire	Phase de mise en œuvre documentaire	Phase de mise en œuvre non documentaire
Localisation-accès	Chercher dans des annuaires la localisation d'un magasin de disques	Demander à des personnes la localisation d'un magasin de disque	Télécharger le disque en format numérique (document numérique)	Acheter le disque et l'emmener avec soi
Sélection-décision	Rassembler les différents catalogues rassemblant les voyages existants	Faire le tour de ses connaissances pertinentes en les interrogeant sur les voyages existants	Faire un tableau comparant les différents types de voyages possibles et permettant de conduire à la décision	Prendre la décision organisant ses arguments dans un dialogue intérieur ou avec des tiers
Synthèse-conception	Rechercher des documents permettant d'imaginer un nouveau voyage	Discuter et faire des visites pour imaginer un nouveau voyage	Rédiger une proposition de voyage inédite en documentant les étapes et les options prévues	Réaliser le voyage selon les étapes et options prévues



Présentation des systèmes d'organisation des connaissances (SOC/KOS)

Les classifications épistémiques universelles de la bibliothéconomie (1)

- Les schémas de classification : la classification décimale de M. Dewey (CDD) et la classification décimale universelle (CDU) de P. Otlet et H. La Fontaine
 - approche épistémologique visant à classer l'ensemble du savoir humain selon une division hiérarchique.
 - faciliter le rangement des exemplaires physiques et fournir une organisation systématique des ouvrages permettant « *au chercheur de repérer des documents pertinents qu'il ne connaît pas encore* » (Hudon 2001).
- CDD classification unique de chaque ouvrage
- CDU combinaison d'indices qui, bien qu'accroissant la précision, n'en facilite pas toujours l'usage.

Les classifications épistémiques universelles de la bibliothéconomie (2)

- Dans un effort similaire de couverture du sens du sujet d'un livre par combinaison d'indices, le bibliothécaire indien S. R. Ranganathan propose en 1924 la « Colon Classification » (CC) basée sur le principe de l'addition de facettes classificatoires (Maniez 1999).
- Chaque sujet doit être qualifié de cinq manières, selon la personnalité, la matière, l'énergie, l'espace et le temps.
- Plusieurs améliorations concernant notamment le caractère plus ou moins universel des facettes retenues dans chaque domaine de connaissance.
- Pour chaque domaine, le principe est toujours de définir un vocabulaire universellement accepté pour faciliter le rangement des livres.

Les thésaurus (1)

- Selon la norme internationale ISO 2788 (1986), les thésaurus nés dans les années 50, sont le « *vocabulaire d'un langage d'indexation contrôlé organisé formellement de façon à expliciter les relations a priori entre les notions (par exemple relation générique-spécifique)* ».
- Langage d'indexation « *ensemble contrôlé de termes choisis dans une langue naturelle et utilisés pour représenter sous forme condensée, le contenu des documents* » (Saadani L. & Bertrand-Gastaldy S. 2000).
 - Descripteurs et non descripteurs (termes interdits), définitions, notes d'application pratique et structure classificatoire : relation d'équivalence intra linguistique (synonymie), relation d'équivalence inter linguistique (traduction), relation hiérarchique, relation d'association.
- CDD & CDU & CC : un indice simple ou composé
 - CC combinaison d'indices élémentaires assemblés selon une syntaxe précise
- Thésaurus : autant de descripteurs que nécessaire

Les thésaurus (2)

- Alors que les classifications organisent les sujets des documents (vedette matière ou subject headings), les termes des thésaurus visent à décrire des concepts.
- La distinction entre sujet et concept est assimilable à la distinction entre parole et langue (Maniez 1999)
 - Les sujets sont en nombre potentiellement infini, les concepts correspondent à un ensemble restreint de notions associées aux ressources cognitives d'une collectivité et dépendant notamment de sa langue (« *Ce qui distingue le concept du sujet est son statut sociolinguistique et son statut cognitif* » Maniez 1999).
 - Le descripteur peut être considéré « toute choses étant égales par ailleurs » comme le meilleur représentant du concept visé (justification des équivalences inter-linguistiques - traduction).
- Concepts du thésaurus vs concepts de l'IA et des ontologies
 - concepts du thésaurus définis à fin d'indexation à partir d'un fonds documentaire pour en faciliter l'interrogation ultérieure.
 - au moins pour partie dépendant des langues et des mises en discours (différence avec les ontologies formelles).

Ontologies formelles et le Web Sémantique

- Liées à l'informatique
 - Continuité de nombreux travaux sur la représentation des connaissances : réseaux sémantiques, cartes conceptuelles, graphes conceptuels
- Popularité qui a principalement bénéficiée du développement du Web Sémantique,
 - Vision prospective et normative du web proposée par Tim Berners Lee en 1994, sans avoir connu, à ce jour, le succès escompté.
- Concepts appréhendés comme des représentations mentales plus ou moins universelles ou comme des catégories a priori largement partagées dans la droite ligne de la philosophie de la connaissance (Guarino 1998).
 - On distingue des ontologies de différents niveaux de généralité : de haut niveau (l'espace, le temps, la matière, les objets, les événements, les actions...) de domaine (médecine, architecture, mécanique..) ; de tâche (diagnostiquer, enseigner), etc.

Caractéristiques des ontologies (1)

- Pas une vocation exclusivement documentaire au sens de l'indexation et de la recherche d'information
 - *Elles visent aussi à participer de l'ingénierie des connaissances d'un domaine* et en particulier à « spécifier explicitement une conceptualisation » pour reprendre les termes de T. Gruber (1993).
- *Egalement conçues à partir d'autres sources d'information* que documentaire
 - Entretiens auprès d'experts, l'analyse de bases de données, ou des conceptualisations ad hoc.
- *L'information dont elles visent à faciliter l'accès est d'abord celle du web invisible,*
 - Les bases de données qui consignent l'information structurée des processus d'affaires et des références techniques.

Caractéristiques des ontologies (2)

- Elles sont le plus souvent conçues pour être *exploitées par des programmes informatiques* (des agents de recherche automatique sur le web)
 - l'utilisateur interagissant avec l'agent à l'aide d'un formulaire ou d'un autre type de langage de requête.
- Par conséquent elles sont *représentées à l'aide de langages formels* :
 - Le standard du W3C (World Wide Web Consortium) est OWL (Ontology Web Language) basé sur RDF (Ressource Description Framework) lui-même exprimé à l'aide de balises XML comme tous les langages du web sémantique.
 - Les classifications exprimées en OWL s'appuient sur une stricte séparation classe/instance, l'héritage de propriétés, l'expression de contraintes de cardinalité et de contraintes logiques sur les relations entres propriétés, etc.

Caractéristiques des ontologies (3)

- *La sémantique des ontologies est une sémantique référentielle*
 - Au sens du positivisme logique, le sens des termes est liée à un référent externe vrai ou faux
- Vision du concept incompatible avec les épistémologies de la philosophie pragmatique (J. Dewey, C. S. Pierce) ou de la tradition herméneutique (répandue dans les sciences humaines et sociales) que nous défendons dans le cadre du web socio sémantique
 - **sens** à l'intersection d'approche logiques/componentielle, contextuelles (lié au corpus), situationnelles ou pragmatiques (ne se réduisant pas aux choses mais étant lié aux situations d'interaction)

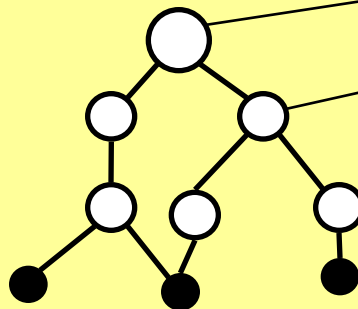
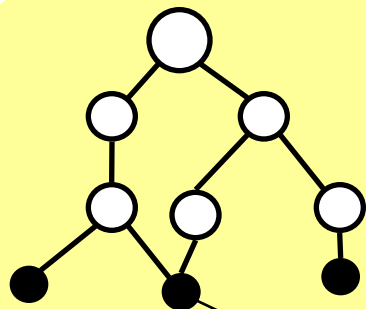
Ontologies sémiotiques et web socio sémantique (1)

- Issus de préoccupations conjointes à l'Ingénierie des Connaissances, au CSCW (Computer Supported Cooperative Work) et au Social Informatics (Turner 2007)
- S'oppose à la vision logiciste du web sémantique
- Basé sur un format de représentation de l'information, la métasémiotique HyperTopic, permettant construire et de partager aisément des ontologies sémiotiques de type cartes de thèmes ou réseaux de description.
 - Même structure hiérarchique que les thésaurus rassemblant des expressions significatives du domaine selon une relation général / spécifique, sans imposer un formalisme logique ou « orienté objet » (pas de relation d'héritage au sens strict, par exemple).

HyperTopic : trois approches de l'item

Thématisation
heuristique (liens
hypertextes)

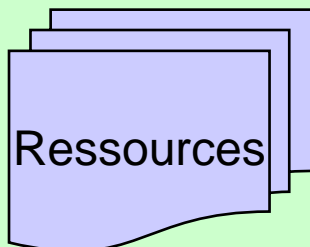
**Ontologies sémiotiques
(attributs heuristiques)**



Point de vue

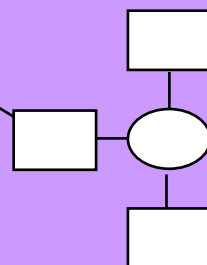
Thème

Ressources



Item

Attributs standards



Documentation de
l'item (fouille de
texte)

Spécification
référentielle :
objets et concepts
(requêtes
logiques)

Ontologies sémiotiques et web socio sémantique (2)

- Son considérées comme des ontologies : elles visent à classifier des situations, des personnes ou des artefacts non nécessairement entièrement documentés.
- Chaque point de vue est en principe défendu par un ou plusieurs acteurs et peut être socialement et/ou cognitivement conflictuel avec un autre – la définition des points de vue est redéfinie dans chaque domaine (différence avec les classification à facette).
- L'organisation de la diversité des points de vue permet de médiatiser la coopération entre des acteurs ou des communautés d'acteurs hétérogènes dans des domaines en partie controversés ou sujets à des interprétations ou des formes d'expériences contrastées.
- Elles doivent donc être évolutives et pouvoir être facilement actualisées

Annuaire de ressources internet collaboratifs et folksonomies (1)

- Annuaire internet et folksonomies mettent en synergie une communauté d'indexeurs coopérant à travers le web dans une logique proche de celle du web socio sémantique.
 - Les annuaire s'inscrivent dans la logique des schémas de classification sans revendiquer l'organisation disciplinaire à laquelle ceux-ci se conforment (Hudon 2001).
- Les annuaire de ressources participatifs pris en charge par des communautés de bénévoles comme l'annuaire « libre » dmoz (<http://dmoz.org/>), étudié par C. Lejeune (2006, 2004).
- Selon leur niveau de réputation, les membres de la communauté virtuelle contrôlent des niveaux plus ou moins haut de la classification,
 - rajouter de nouvelle branches dans les domaines au sein desquels ils font « autorité », supprimer les descriptions effectuées dans la « notice » par des participant moins renommés, etc.

Annuaire de ressources internet collaboratives et folksonomies (2)

- Succès actuel des folksonomies dans le contexte des business model du Web 2.0. : l'utilisateur indexe lui-même les documents dont il est souvent le promoteur (Del.icio.us <http://del.icio.us> ou Flickr <http://www.flickr.com>)
- Faible cohérence des descripteurs (synonymie, polysémie, non explicitation des facettes prises en compte, absence de relation sémantique...) mais réel succès.
- Pratiques d'indexation sociales, (Ertzscheid et Gallezot 2006), 3 raisons du succès :
 - Faible effort cognitif requis par leur utilisation en comparaisons des classifications épistémiques de la bibliothéconomie et, d'autre part
 - Fonction de régulation offerte par la mise en visibilité des mots-clefs déposés par l'ensemble des utilisateurs qui permet d'avoir un effet de feed-back rapide sur leur popularité et leur degré de couverture (Ertzscheid & Gallezot 2006).
 - Possibilité d'accès quasi immédiat à la ressource pour désambigüiser

Annuaire de ressources internet collaboratifs et folksonomies (3)

- Pas d'innovations du point de vue de l'organisation conceptuelle des descripteurs.
- L'innovation se situe dans le processus collaboratif de construction des descripteurs et dans le processus d'indexation associé à cette construction à partir d'un flux de documents primaires très hétérogènes et dont le volume s'accroît très rapidement.
- Hybridation souhaitable entre des dispositifs professionnels de type schémas de classification, thésaurus et ontologies et ces nouveaux dispositifs de gestion collaborative de l'information numérique -> Web Socio Sémantique



Comparaison des SOC/KOS selon sept critères

Evaluation des SOC du point de vue de la ROI

- Les critères que nous passerons en revue sont les suivants :
 1. le degré de formalité de la métasémiotique ou de la métalangue utilisée,
 2. le caractère théorique ou a-théorique de la classification sous-jacente,
 3. le type de communauté responsable de la conception,
 4. la nature des sources utilisées pour fonder la classification,
 5. les théories de la signification sous-jacentes aux termes,
 6. les modalités de mise à jour
 7. les systèmes de consultation offerts.
- Mise en relation avec le caractère documentaire ou extra documentaire de l'enquête et la typologie de celle-ci (localisationaccès, sélection-décision, synthèse conception)

Degré de formalisation du langage et des combinaisons syntaxiques

Pas de syntaxe pour la gestion combinée des termes	Méta sémiotique explicite	Méta langage formel
<ul style="list-style-type: none">• Classification décimale de Dewey• Folksonomie	<ul style="list-style-type: none">• CDU• Thésaurus• Ontologie sémiotique• Annuaire collaboratif internet• Carte conceptuelle	<ul style="list-style-type: none">• Ontologie formelle

- Plutôt la cause d'autres phénomènes que fournissant une fonctionnalité d'usage directe

Présence ou non de justifications à caractère théorique

Pas de théorie explicite du domaine	Théorie faisant l'objet d'une controverse explicite	Théorie disciplinaire (souvent positiviste)
<ul style="list-style-type: none">•Folksonomie•Annuaire de ressources internet•Cartes conceptuelles (point de vue subjectif revendiqué)	<ul style="list-style-type: none">•Ontologies sémiotiques (points de vue complémentaires mais potentiellement conflictuels)	<ul style="list-style-type: none">•Classifications universelles (CDD, CDU...)•Thésaurus•Ontologies formelles

- Très utile si la recherche s'inscrit dans un cadre théorique compatible (divisions disciplinaires), handicapant dans le cas contraire
- Mieux vaut alors s'adapter à des théories ad hoc (ontologies sémiotiques) ou pas de théorie (interdisciplinarité de la problématique...)

Type de communauté impliquée dans la conception des SOC

Conception par des professionnels des bibliothèques ou de la documentation	Conception par des ingénieurs de la connaissance (avec des experts du domaine)	Conception participative régulée (expertise multiple basée p.e. sur la réputation)	Conception participative grand public non contrôlée (sagesse des foules ou effet de mode)
<ul style="list-style-type: none"> • Classifications universelles • Thésaurus 	<ul style="list-style-type: none"> • Ontologies formelles • Cartes conceptuelles 	<ul style="list-style-type: none"> • Annuaire internet • Ontologies sémiotiques (éventuellement avec la médiation d'un ingénieur de la connaissance) 	<ul style="list-style-type: none"> • Folksonomies

- Plus les enquêtes sont complexes plus l'information documentaire est l'enjeu de controverses, plus il faut articuler des points de vue divergents :
- Participation des experts et usagers notamment nécessaire dans les enquêtes de sélection-décision et synthèse-conception » qui articulent des points de vue très différents

Nature des sources utilisées

Basé sur un fonds bibliographique	Structuration du domaine empruntant à d'autres sources d'information
<ul style="list-style-type: none">• Classifications universelles• Thésaurus• Folksonomies• Annuaire internet	<ul style="list-style-type: none">• Cartes conceptuelles• Ontologies sémiotiques• Ontologies formelles

- Important quand la recherche d'information est également extra-documentaire !

Théorie de la signification sous-jacente

Sens basé sur une sémantique référentielle et possibilité d'accès selon cette sémantique	Sens rhétorico-herméneutique et accès selon les associations « heuristiques »
<ul style="list-style-type: none">• Classification universelle (CDD)• Ontologies formelles	<ul style="list-style-type: none">• CDU• Cartes conceptuelles• Annuaire internet• Thésaurus• Ontologie sémiotique• Folksonomie

- Essentiel pour les enquêtes de localisation accès (ou sélection décision basée sur critères mesurables)

- Préférable quand les enquêtes de type sélection-décision font intervenir des critères heuristiques et subjectifs ou dans les enquêtes de type synthèse-conception (multipoints de vue)

Fréquence et facilité de mise à jour

Mise à jour rare et complexe (maintien de la cohérence)	Mise à jour systématique régulée par l'évolution du fonds	Mise à jour progressive et négociée	Mise à jour fréquente, facile, immédiate
<ul style="list-style-type: none">• Classification universelles• Ontologies formelles	<ul style="list-style-type: none">• Thésaurus	<ul style="list-style-type: none">• Annuaire internet• Ontologie sémiotique• Cartes conceptuelles	<ul style="list-style-type: none">• Folksonomie

- 
- Plus la connaissance évolue rapidement plus l'évolution des SOC doit être facilitée...

Systemes de consultation

De type formulaire ou lié à l'interopérabilité entre programmes	Vue dépliée permettant l'interprétation des descripteurs	Systemes de navigation hypertexte ergonomiques
<ul style="list-style-type: none">• Ontologies formelles	<ul style="list-style-type: none">• Classifications universelles• Thésaurus	<ul style="list-style-type: none">• Ontologies sémiotiques• Cartes conceptuelles• Folksonomie• Annuaire internet

- Utilisation à travers des formulaires ou par des programmes informatiques

- Vue dépliée permettant de préciser la sémantique des termes lors de la navigation (browsing)



Conclusion

- Une bonne compréhension des différents systèmes d'organisation des connaissances existants doit permettre d'imaginer des solutions hybrides d'aide à la recherche ouverte d'information dans le contexte
 - de la numérisation accélérée des ressources documentaires
 - de la participation de plus en plus aisée des experts et des utilisateurs.
- RIO implique de penser la complémentarité des SOC avec l'usage des moteurs de recherche
- Les potentialités de l'indexation interprétative (manuelle) sont très importantes et doivent constituer une voie de recherche majeure à l'intersection des sciences de l'information, des sciences de la communication, de l'ingénierie des connaissances et du CSCW.