Unleashing the power of data through organization

Structure and connections for meaning, learning, and discovery

Dagobert Soergel

Department of Library and Information Studies
Graduate School of Education
University at Buffalo

ISKO UK 2015

See the full paper for detail and references

The Future of Knowledge Organization

Knowledge organization is needed everywhere

Create the future of KO

Think BIG. Think answers not pointers.
Focus on substantive data

Many areas, tasks, and functions
that could profit from KO principles

Engage with Ontologies, AI, data modeling

Areas, tasks, and functions

1. Knowledge bases for question-answering and cognitive systems
2. Knowledge base for information extraction from text or multimedia
3. Linked data
4. Big data and data analytics. Data interoperability and reuse
5. Interoperability of operational information systems.
   Electronic health records (EHR) as an example
6. Information systems in the enterprise
7. Influence diagrams (causal maps), dynamic system models,
   process diagrams, concept maps, and other node-link diagrams
8. Knowledge organization for understanding and learning
9. Knowledge transfer between domains

Unification

- **across applications**
- **across types of data** (example: organization database treated like classification)
- **across disciplines**, supports knowledge transfer from one discipline domain to another
- **across languages** (precise definitions)
- **across cultures**, across organizations (organizational cultures)
- **across worldviews**

Part 2
The application of Knowledge Organization

2.1 Knowledge bases for question-answering and cognitive computing

Knowledge base	Some KOS used
CYC Common sense knowledge	CYC Ontology, including entity types, relationship types, and entity values
IBM Watson Custom KB for applications	An extensible inventory of relationship types
Google Knowledge Graph Huge database of varied kind of data (Starr 2014)	schema.org for entity types and relationship types
DBpedia Large database of statements extracted from Wikipedia	DBpedia Ontology (E-R schema) Authority lists for individual entity values (instances), each identified by a URI.
GDELT Event reports	CAMEO Coding Scheme for events Own list of 300 themes, World Bank Taxonomy themes 2,300 emotions and themes (from 24 sentiment analysis packages) US government geonames standards

2.2 Knowledge base for information extraction from text or multimedia

Often only text is considered, but information can be extracted from graphs and video (for example, identifying people by face recognition and relationships between people from analyzing scenes). In the following text+

Information extraction

- **Entity extraction** (Named-entity recognition)
  Locating references to entities in text+, associate with a unique identifier.

- **Information extraction**
  Formally represent the propositions the text makes about these entities.

Information extraction both uses and feeds knowledge bases for question answering.

KOS for information extraction

Information extraction needs much knowledge, which must be properly organized into KOS

- **Linguistic knowledge**: morphological, part-of-speech, and lexical (meaning). Lexicalized phrases.
- **Large KOS listing entity values** and their (multiple) names (persons, organizations, places, concepts/subjects, ...)
- **Knowledge supporting word sense disambiguation** (WSD). Both linguistic knowledge and world knowledge.

2.3 Linked data

- **Entity-relationship data model**
- **Data from independent data sets can linked**
- **Key implementation component of the Semantic Web**
- **Enormous opportunity for KO**.
  - Deploying KOS data on the Web and have them more widely used.
  - Linked data require properly structured and often very large KOS.

Linked data

- **The more pervasive standardization** with respect to
  - entity types
  - relationship types
  - entity values

  the more successful linked data searching will be

- **This is a problem of knowledge organization**

Drug <*hasName*> Text

Drug <*hasGenericVersion*> Drug

Drug <*hasActiveIngredient*> ChemicalSubstance

Drug <*hasClinicalPharmacologyDescr*> Text

Drug <*hasIndicationDescr*> Text

Drug <*hasContraIndicationDescription*> Text

Drug <*administeredVia*> RouteOfAdministration

DBDrug <*hasName*> Text

DBDrug <has*GenericName*> Text

DBDrug <has*CASRegistryNumber*> URI

DBDrug <*hasAbsorptionDescr*> Text

DBDrug <*hasBioTransformDescr*> Text

DBDrug <*hasPharmacolDescr*> Text

DBDrug <*hasProteinBindRate*> Pct

DBDrug <*hasIndicationDescr*> Text

DBDrug <*hasPossibleDiseaseTarget*> Disease

DBDrug <*hasContraIndicationInsert*> Document

DBDrug <*hasDosageForm*> DosageForm

2.4 Big data and data analytics. Data interoperability and reuse

Example 1. Merging like datasets

- **Research question**: Factors affecting school success
- **Need large sample, so merge data sets** with anonymized data on individual students and test scores from many US states (many European countries)
- **Problem**: this **works only if variables are defined the same way in all data sets**
  – Factors such as *socio-economic status of the student* or *home environment*
  – Concepts and skills covered in the tests.
- **This is a knowledge organization problem**

Example 2. Linking datasets

- **Research question:** relationships between
  · *per capita income*,
  · *how people feel about the economy*, and
  · *birth rate*
  Unit of analysis: Locality

- **The variables needed are in three different data sets**:
  1 per-capita income by locality
  2 Twitter messages (analyze for sentiment)
  3 Birth rate by locality
  The data sets need to be linked so that for each locality we have values for the three variables

- **Problem:** The **ability to link these data sets depends on the linking variable, locality, being defined the same way** and identifiable (a problem with Twitter)

2.5 Interoperability of operational information systems
Electronic health records (EHR) as an example

- Interoperability of EHR data is an **obvious must**, but far from solved.
- **Needs KOS for**
  - race/ethnicity, age, sex
  - bodily or mental functions or conditions
  - diseases
  - medical procedures
  - drugs
- **Worked on heavily, mainly by people in biomedical informatics / biomedical ontologies.**
- Given here as one example of the importance of KO for operational systems.

2.6 Information systems in the enterprise

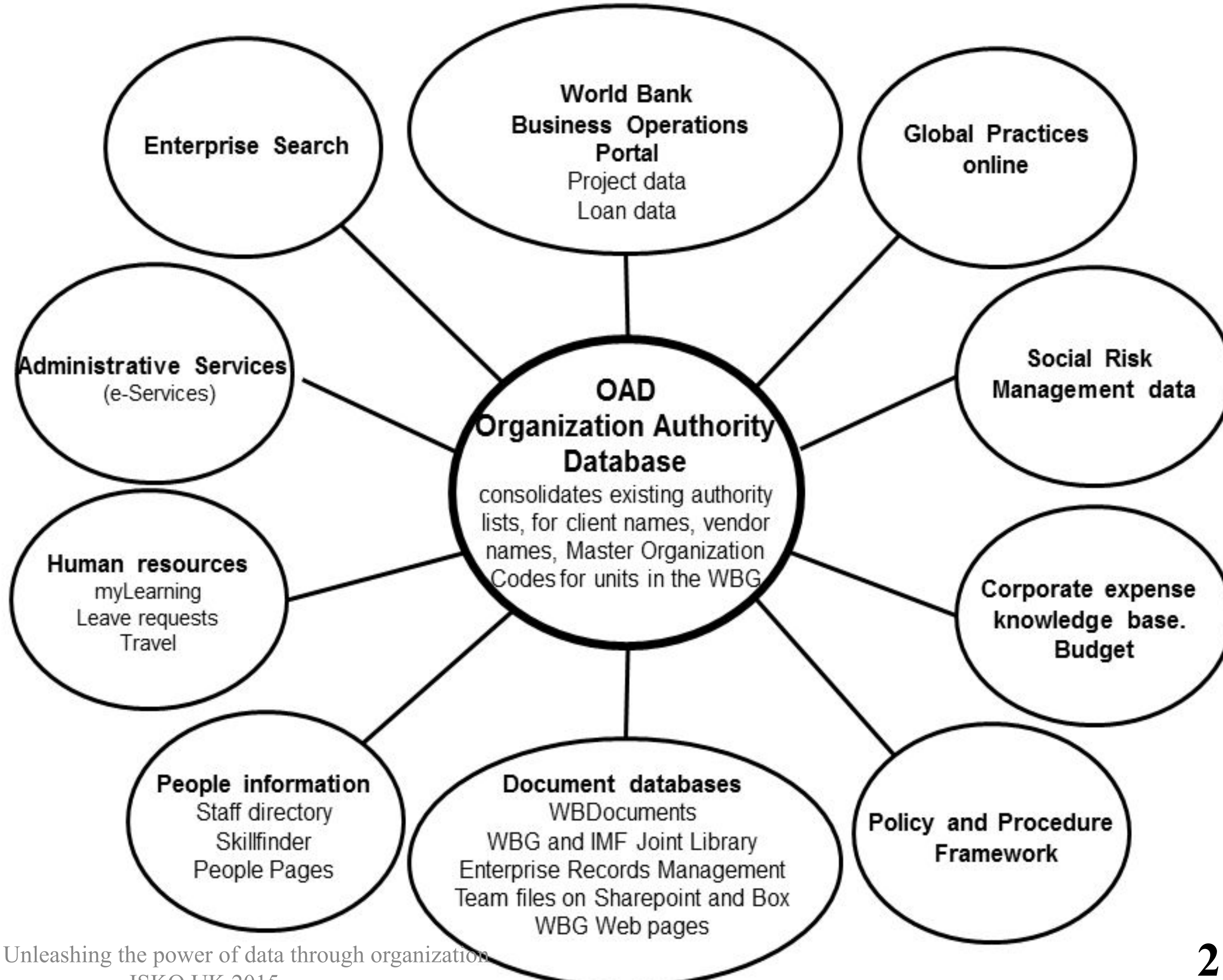
Example 1

- **Problem:**
  Many organizations do not know in a central place what data they have
- **Solution:**
  – **Develop an enterprise-wide entity-relationship conceptual data schema** (an enterprise ontology, an enterprise data model, the modern version of a data dictionary), using ideas from Web standards.
  – Use this to **organize an inventory or registry of all data systems** in the organization and the specific pieces of data in each.

Example 2

Unified authority database for Organizations

considered for the World Bank Group (WBG)


OAD
Organization Authority Database
consolidates existing authority lists, for client names, vendor names, Master Organization Codes for units in the WBG
World Bank Business Operations Portal
Project data
Loan data
Enterprise Search
Global Practices online
Administrative Services
(e-Services)
Social Risk Management data
Human resources
myLearning
Leave requests
Travel
Corporate expense knowledge base. Budget
People information
Staff directory
Skillfinder
People Pages
Document databases
WBDocuments
WBG and IMF Joint Library
Enterprise Records Management
Team files on Sharepoint and Box
WBG Web pages
Policy and Procedure Framework

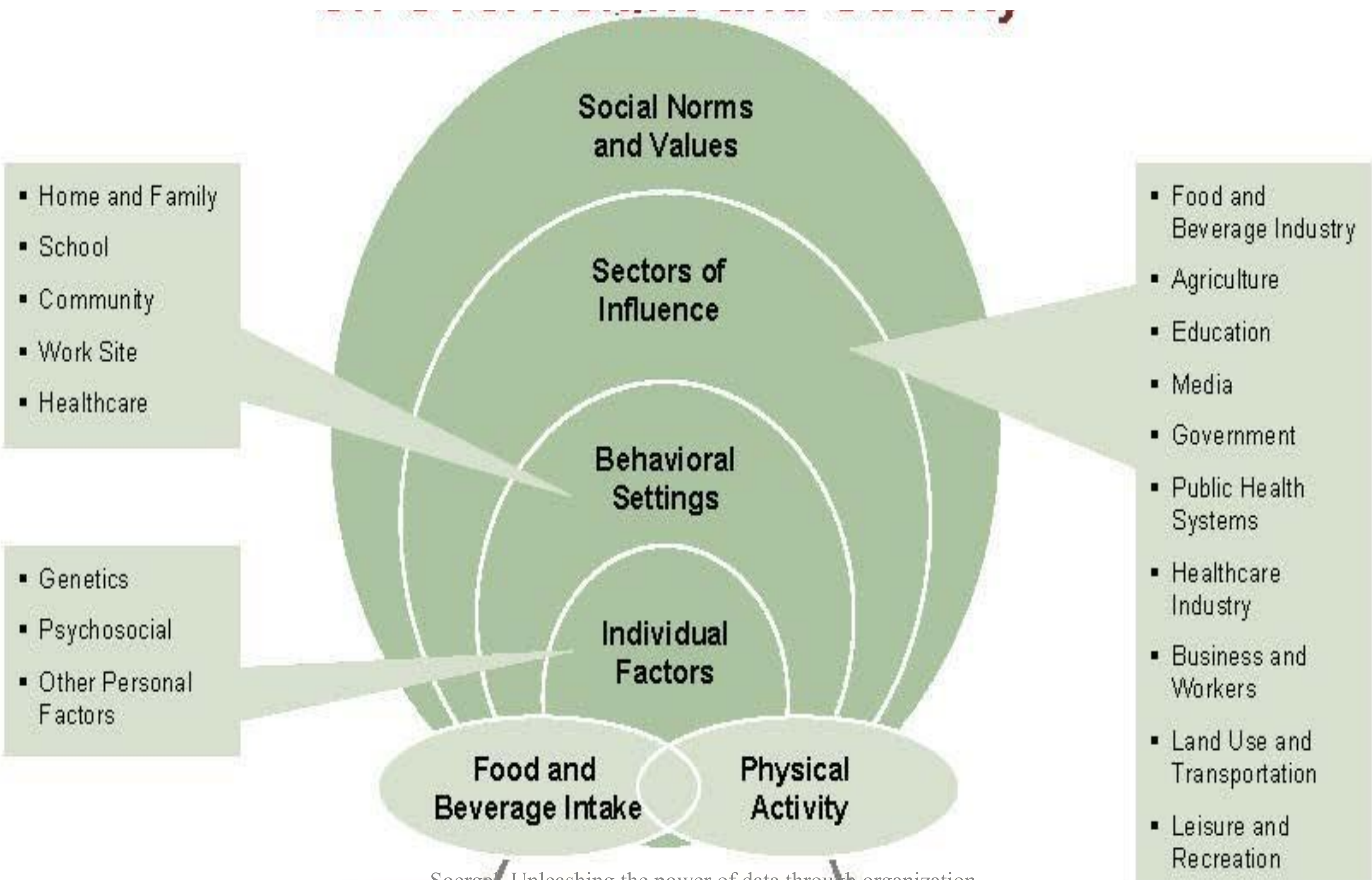
Example 2 cont.

- The enterprise-wide Organization Authority Database should be structured exactly like a hierarchical thesaurus: Just like concepts, the organizations form a hierarchy, and they have multiple names

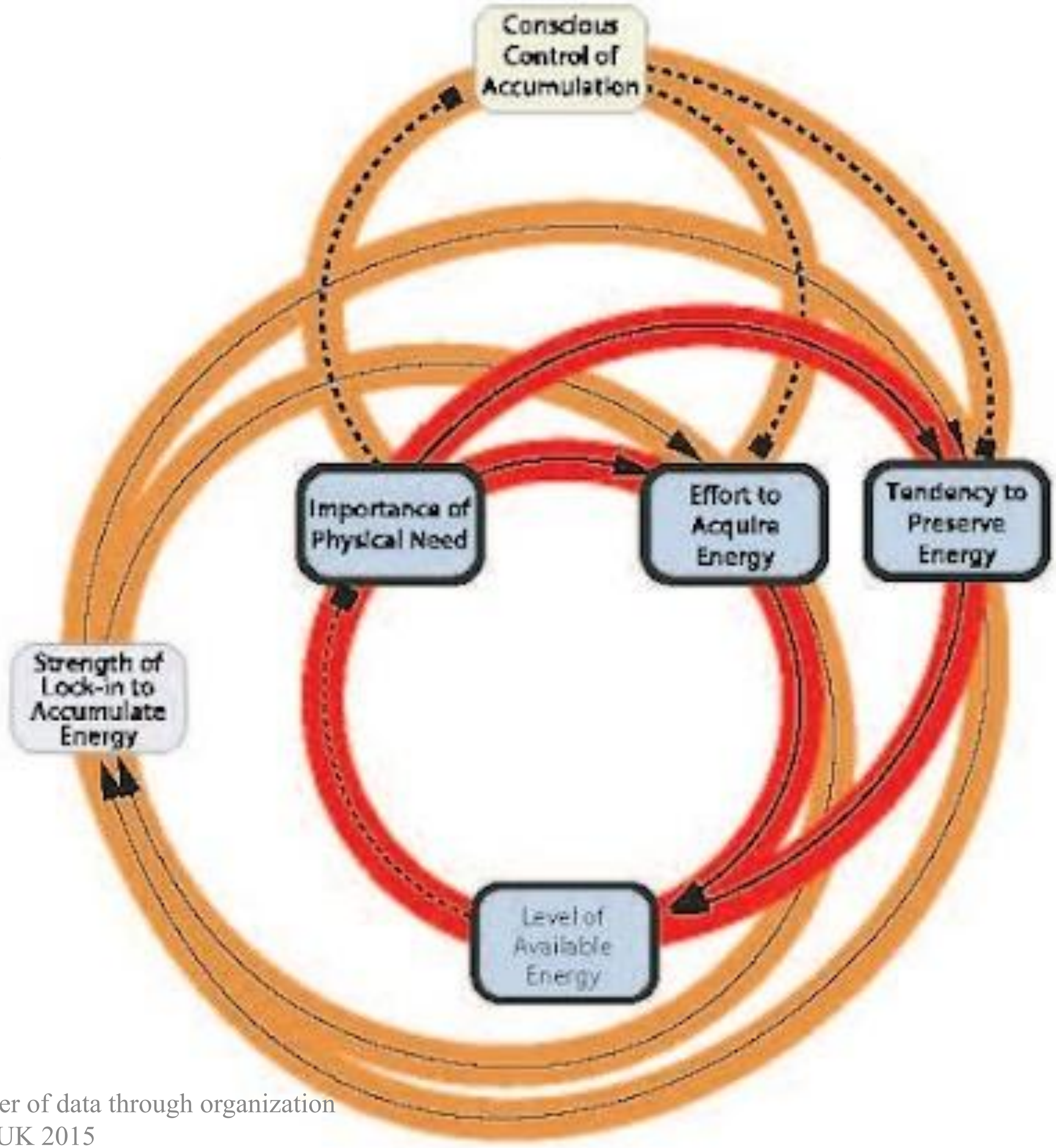
2.7 Node-link diagrams

- Causal maps (influence diagrams)
- Dynamic system models
- Process diagrams
- Concept maps
- Other

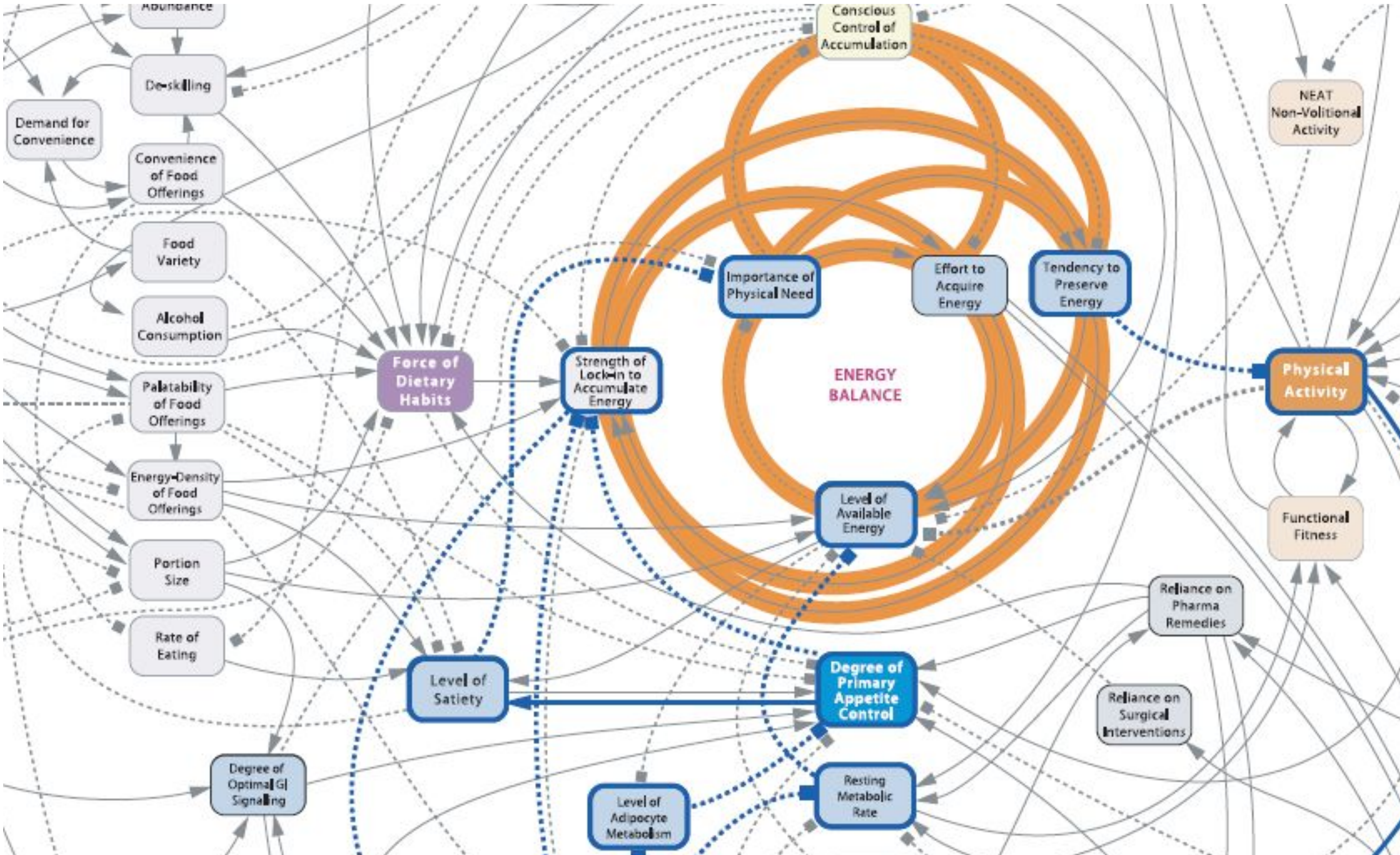
Influences on overweight and obesity

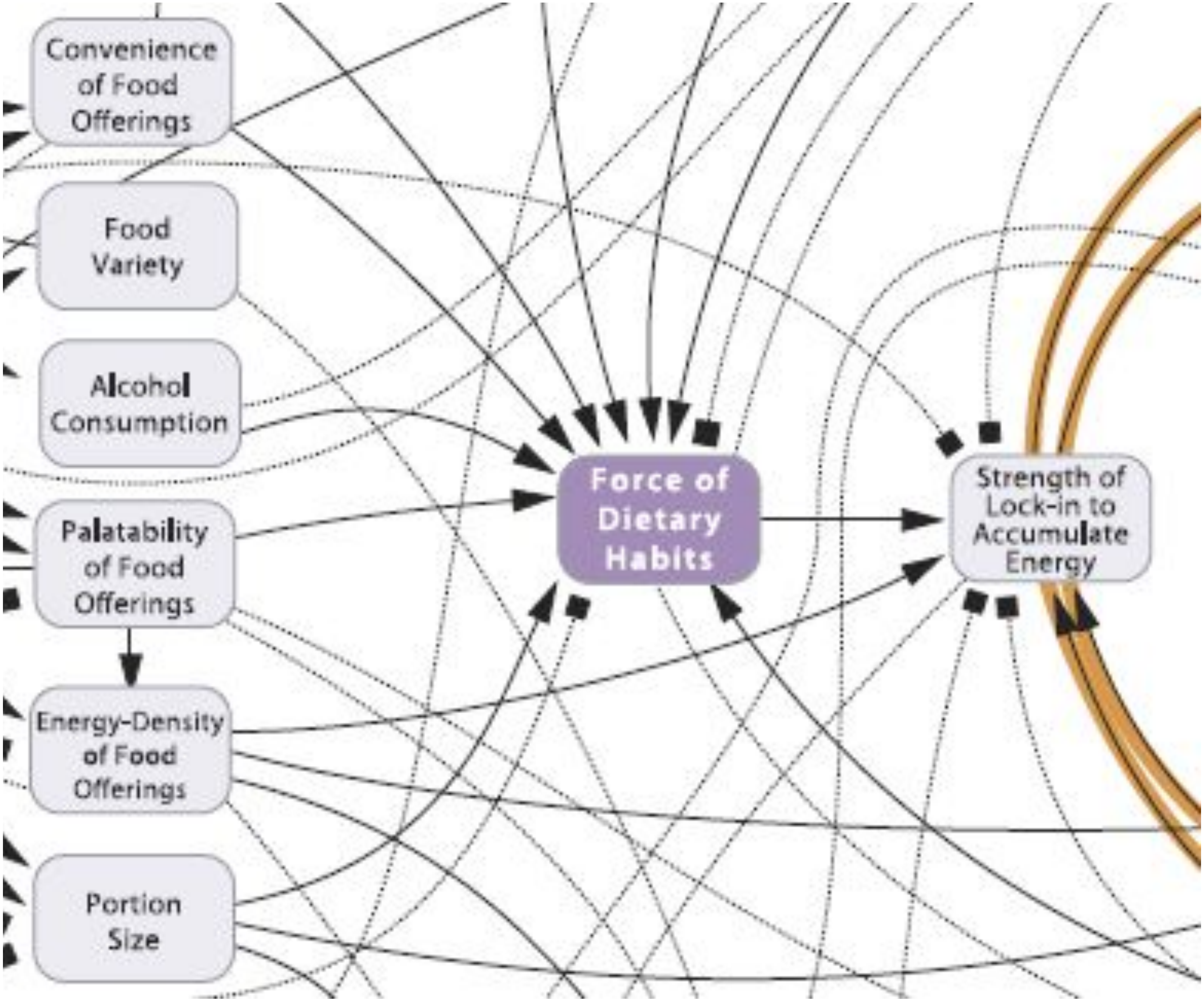


shiftN causal map for obesity

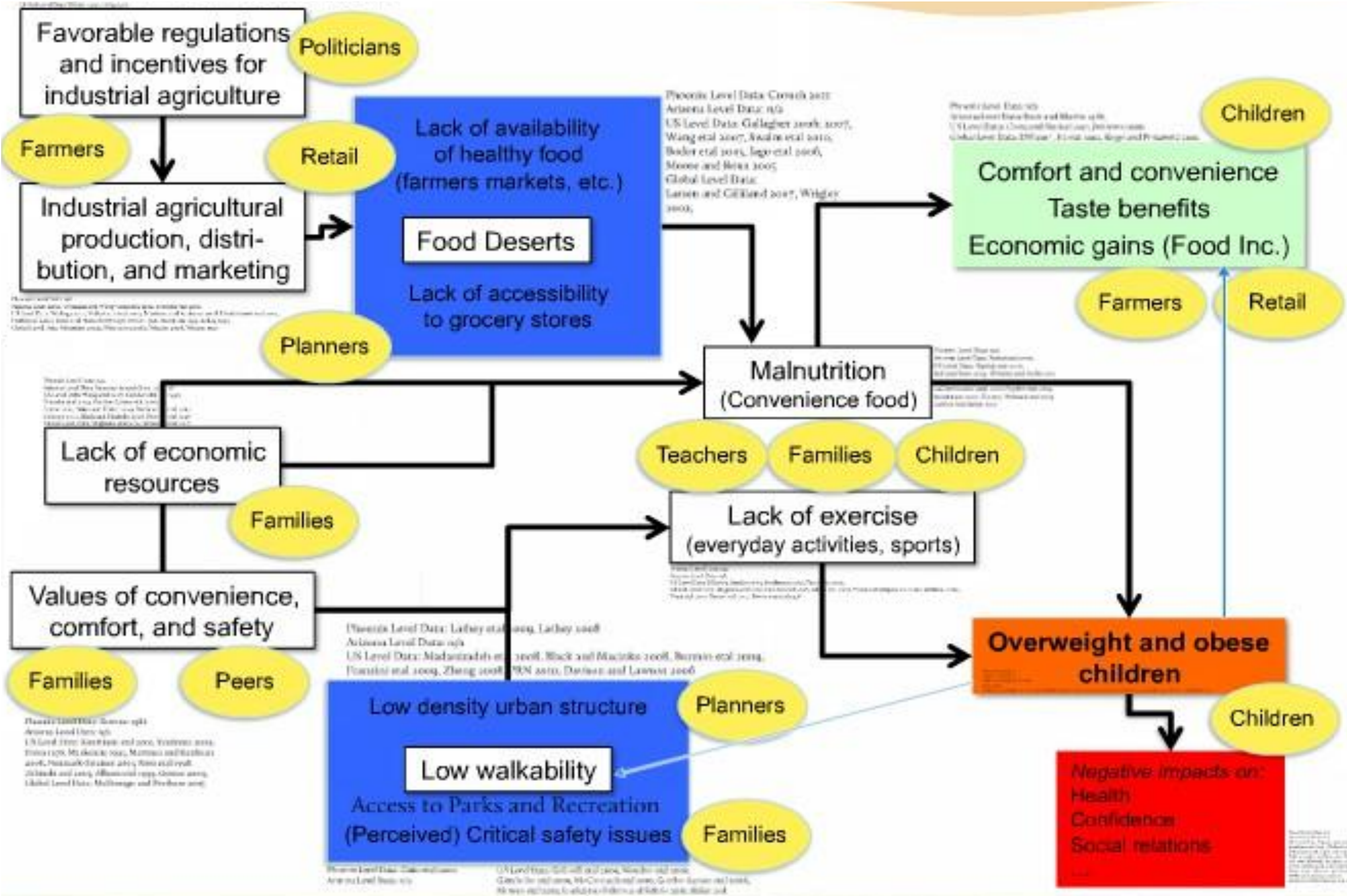


Segment the large and detailed shiftN causal map for obesity





Convenience of Food Offerings
Food Variety
Alcohol Consumption
Palatability of Food Offerings
Energy-Density of Food Offerings
Portion Size
Force of Dietary Habits
Strength of Lock-in to Accumulate Energy


Favorable regulations and incentives for industrial agriculture
Politicians
Farmers
Retail
Industrial agricultural production, distri-bution, and marketing
Lack of availability of healthy food (farmers markets, etc.)
Food Deserts
Lack of accessibility to grocery stores
Planners
Children
Comfort and convenience
Taste benefits
Economic gains (Food Inc.)
Farmers
Retail
Malnutrition (Convenience food)
Teachers
Families
Children
Lack of economic resources
Families
Lack of exercise (everyday activities, sports)
Values of convenience, comfort, and safety
Families
Peers
Overweight and obese children
Children
Low density urban structure
Planners
Low walkability
Access to Parks and Recreation
(Perceived) Critical safety issues
Families
Negative impacts on:
Health
Confidence
Social relations

KO issues

- **Arranging variables in a meaningful order**
- **Mapping variables from one model to another**

Coming up later

- **Merging node-link diagrams**
- **Linking node-link diagrams**

shiftN causal map variables. Top level with example detail (arranged by DS)	
Individual	**Environment**
Engine Energy balance Conscious control of accumulation Effort to acquire energy Strength of lock-in to accumulate energy	
Physiology Degree of primary appetite control by brain Genetic and/or epigenetic predisposition	
Food consumption Force of dietary habits Tendency to graze Demand for convenience Food exposure Food variety	**Food production** Societal pressure to consume Demand for health Pressure to improve access to food offerings Cost of ingredients
Individual physical activity Level of transport activity	**Physical activity environment** Dominance of motorised transport Opportunity for unmotorised transport
Individual psychology Food literacy Stress	**Social psychology** Exposure to food advertising Peer pressure

Some (approximate) matches and non-matches between 4 lists of variables

shiftN	Kaplan	Nanotechnology	Downey' list
Engine			
Energy balance	Energy balance		
	Energy intake		
	Energy expenditure		
Conscious control of accumulation			lack of self-control
Effort to acquire energy			
			Response to food cues
Physiology			
Appetite control by brain			
Genetic & epigenetic predisposition			genetics epigenetic factors
Food consumption	Food and bev. intake		overeating
Force of dietary habits			
		Malnutrition (conv. foods)	high fruct. corn syrup
Food production	Food & bev. industry	Agricultural production	agricultural policies
		Food deserts	food deserts
Cost of ingredients			
Indiv. physical activity	Physical activity	Exercise & physical activity	
		Lack of exercise	Low physical activity

More uses of node-link diagrams

In biology and in industrial engineering

- diagrams of sequential and interrelated processes that lead to some outcome or state

In biology

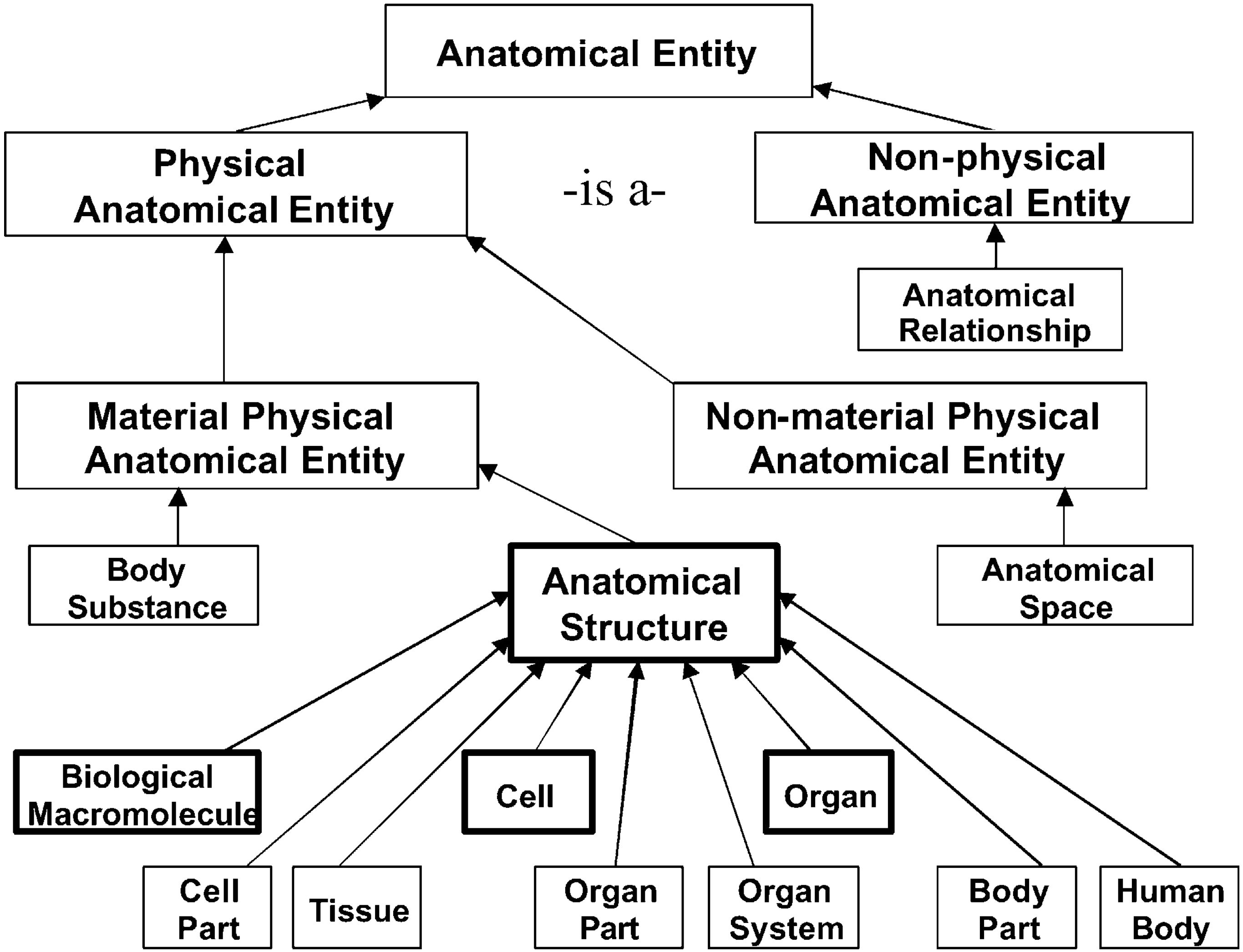
- diagrams of signaling pathways,
- diagrams of metabolic networks,
- diagrams of gene regulatory networks

Concept maps

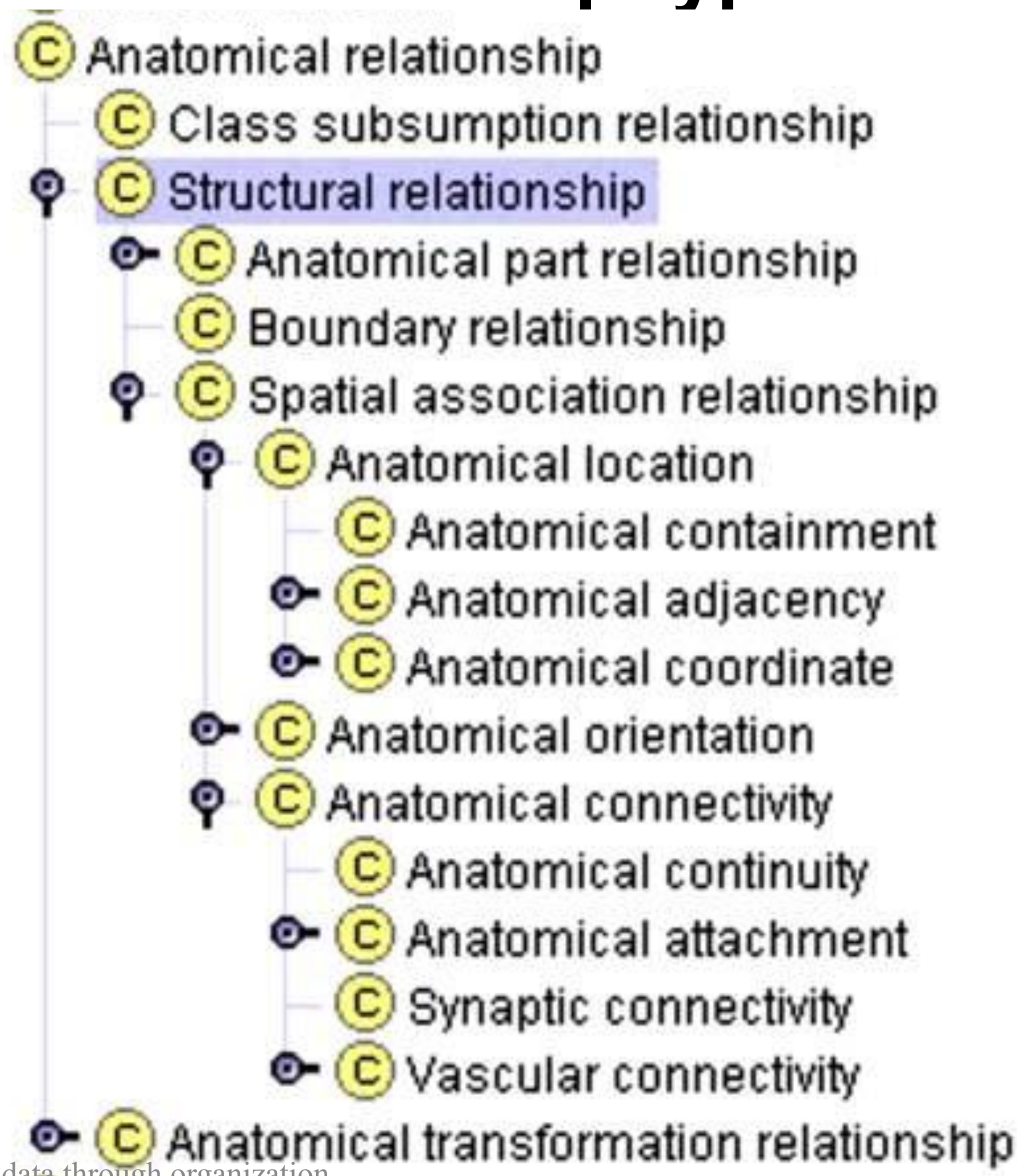
- Used as thesaurus displays since the 1950s
- Resurfaced forcefully in education
- If you know of earlier uses, let me know

2.8 Knowledge organization for understanding and learning

Foundational Model of Anatomy: Entity types



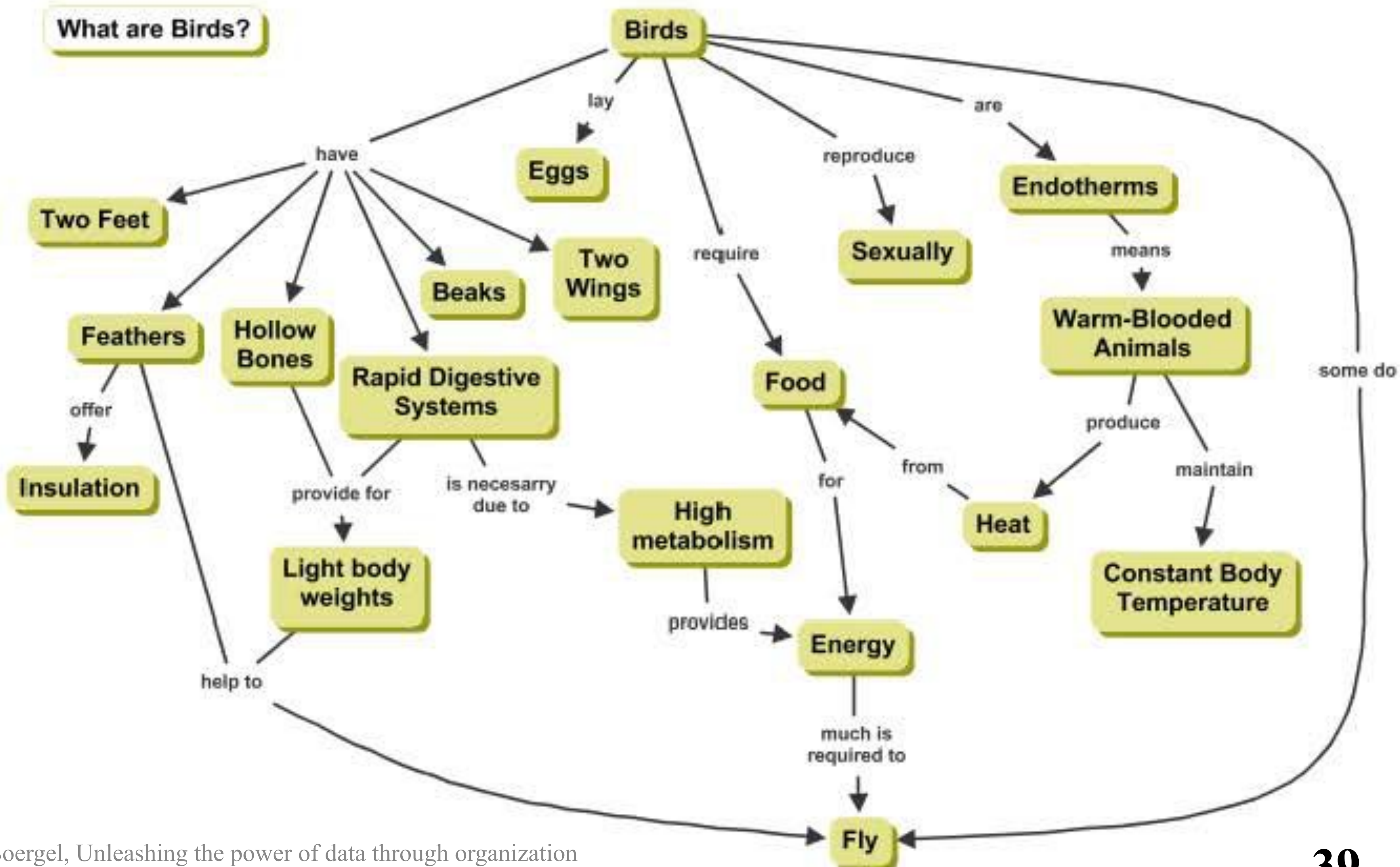
Foundational Model of Anatomy: Relationship types



Hypothesis

Students who are taught anatomy using the
Foundational Model of Anatomy
have a better grasp of the structure of the body.

Concept map about birds



Concept map hypotheses

The bird concept map will allow learners to form a better internal representation of a bird as a system.

Constructing concept maps will help learners to develop a better understanding (a better structured mental model) of the topic.

Meaningful arrangement
Classification of animals

a. **Britannica Elementary: Menu for *Animal Kingdom***

Thoughtless arrangement, devoid of any meaning

One of many such examples from children's websites

b. Meaningful arrangement

Based on the modern cladogram shown in c.

c. Vertebrates cladogram

Britannica Elementary: Menu for *Animal Kingdom*

Thoughtless arrangement, devoid of any meaning

Animal Kingdom: Meaningful arrangement based on modern science

Animals without a spine (*invertebrates*)

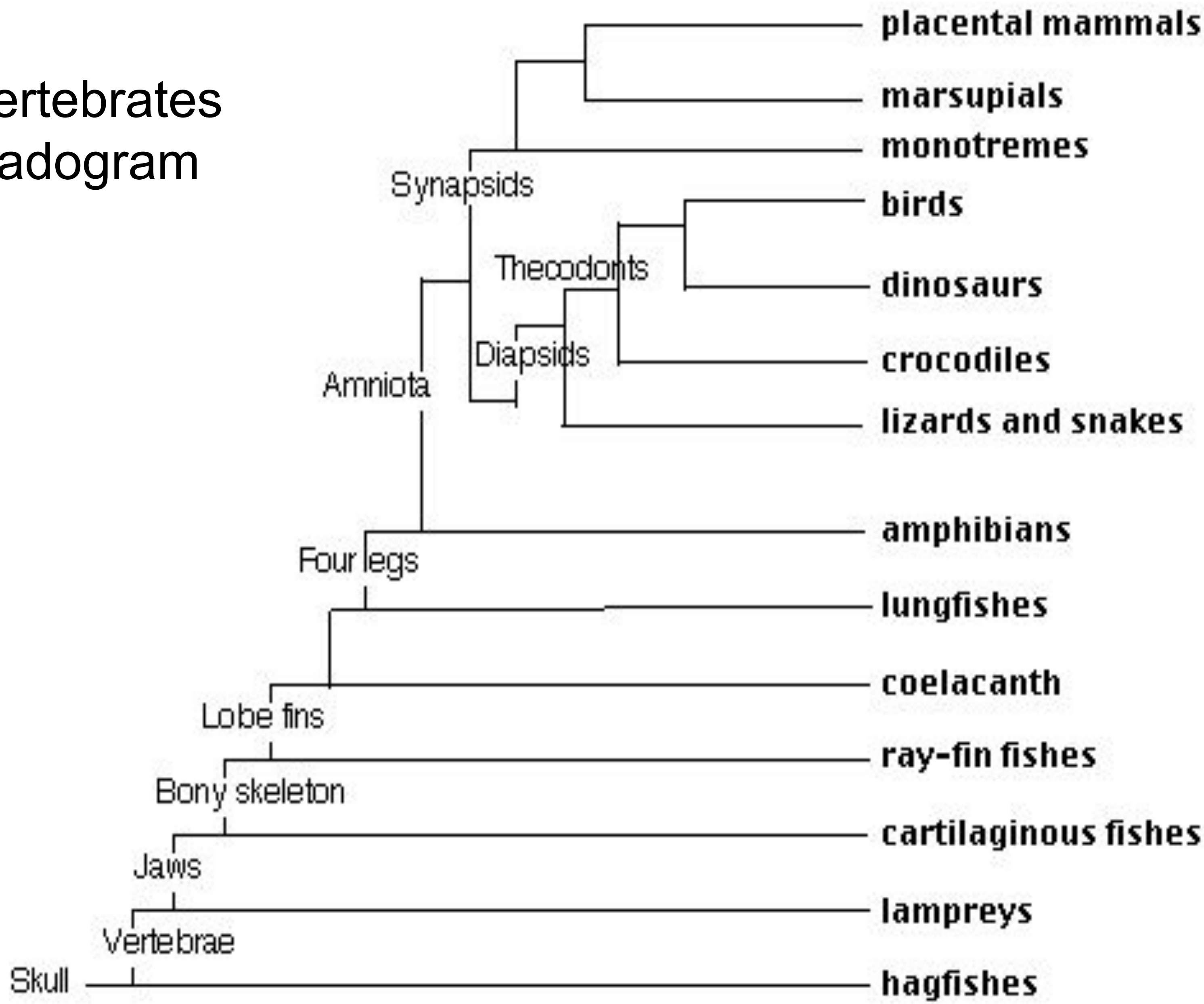
- Snails, octopus, mussels (***mollusks***)
- Bugs (insects), spiders, crabs (***arthropods***)

Animals with a spine (*vertebrates*)

- ***Fish***
- Frogs, toads, salamanders (***amphibians***)
- Lizards&snakes, crocodiles, dinosaurs, birds
  - Lizards&snakes, crocodiles, dinosaurs (***reptiles***)
  - ***Birds***
- Elephants, whales, cows, dogs, bats, mice, monkeys, apes, humans (***mammals***)

Note: Could simplify, add pictures

Vertebrates cladogram



Meaningful arrangement hypothesis

Young students who use the animal home page with the meaningful arrangement will over time absorb the sequence and perceive a progression. When much later in biology the structure of the animal kingdom and the evolution of animals are discussed, these students will understand more quickly.

2.9 Knowledge transfer between domains

Management styles and educational styles compared		
Style of social interaction	**Management style**	**Educational style**
Autocratic, authoritarian, directive	**Autocratic, authoritarian, directive (coercive), top-down**	**Direct instruction, teacher-centered Teacher as formal authority, expert**
Military style	**Military style**	**Military style**
Paternalistic	**Paternalistic**	
Authoritative (visionary)	**Authoritative (visionary)**	
Persuasive	**Persuasive**	
Coaching	**Coaching**	**Teacher as facilitator**
Individual inner discipline, motivation, agreement with norms		**Montessori**

Figure 17. Management styles and educational styles compared		
Style of social interaction	**Management style**	**Educational style**
Participatory, democratic	**Participatory (democratic), consultative**	**Democratic and Free Schools**
Collaborative, teamwork	**Collaborative, teamwork**	**Cooperative Learning Teacher as facilitator, delegator**
Self-directed groups	**Holacracy, self-management in groups**	
Laissez-faire, free-wheeling	**Laissez-faire**	**Open Schools (and Classrooms) (Summerhill)**
Chaotic	**Chaotic**	
People try their own thing		**Inquiry-based learning, student-centered (related to constructivism) Teacher as facilitator, delegator**

Part 3
General observations on knowledge organization and its role

3.1 Better data modeling

- **Entity-relationship modeling is fundamental**
  Kudos to Peter Chen (1976) and precursors

- **Three past blunders**

  1 Attributes as elements in entity-relationship modeling

  2 Calling relationships *properties*, as is done in RDF

  3 Using only binary (two-way) relationships

Part 4
Conclusions

Conclusions 1

- **Many applications of KOS.**
- **Consider both**
  - **requirements for machine processing, specifically inference, and**
  - **requirements for human processing, specifically meaningful arrangements that assists in making sense**

Conclusions 2

- Many opportunities for people with good training in KO to improve KOS now used
- Prepare students for that, specifically
  - Students should have a basic understanding of logic, formal ontology principles, inference, and complex queries
  - Foster the ability to discern meaningful structures and then convey structure and meaning through good document design.
  - Foster the ability to work with researchers on defining variables, determining data collection methods, and curating, and sharing data , all to improve interoperability and reusability.

Conclusions 3

- We need more communication between the following largely separated communities:
  - Knowledge Organization
  - Semantics in linguistics and terminology
  - Knowledge representation in artificial intelligence
  - Ontology
  - Data Modeling
  - Semantic Web

The Future of Knowledge Organization

Knowledge organization is needed everywhere

Create the future of KO

Think BIG. Think answers not pointers.
Focus on substantive data

Many areas, tasks, and functions
that could profit from KO principles

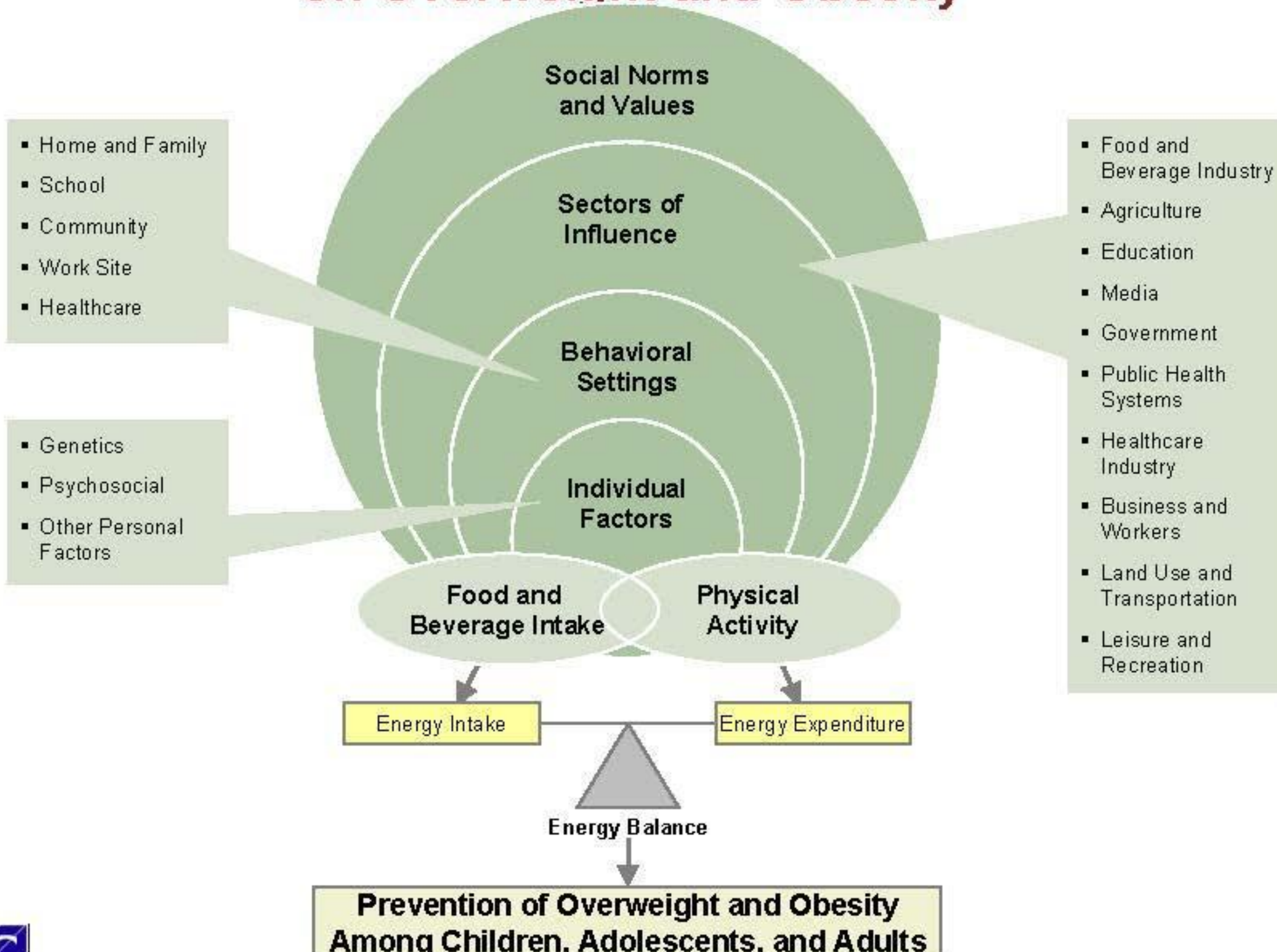
Engage with Ontologies, AI, data modeling

Dagobert Soergel

dsoergel@buffalo.edu

www.dsoergel.com

An Ecological Framework for Organizing Influences on Overweight and Obesity



Adapted from: Koplan JP, Liverman CT, Kraak VI, editors. Preventing childhood obesity: health in the balance. Washington, DC: Institute of Medicine, National Academies Press; 2005.