# IMI共通語彙基盤の目指すところ

武田英明
国立情報学研究所・教授　/　情報処理推進機構・専門員
takeda@nii.ac.jp

# IMI共通語彙とは何か

- 概念辞書
  - 構造をもった概念辞書
  - 構造の転写
  - 典型（プロトタイプ）としての概念

# 語彙に関わる考え方

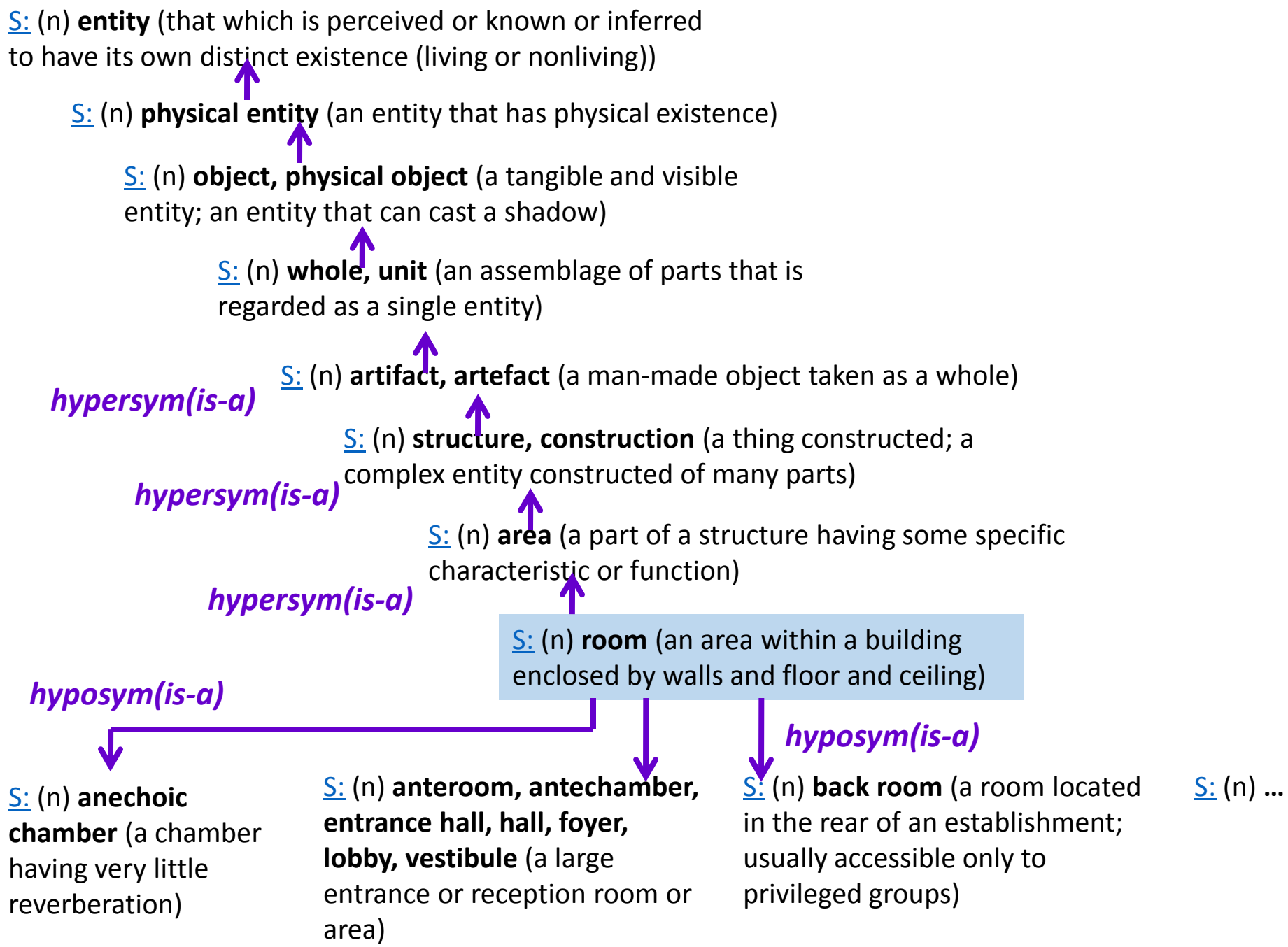
- オントロジー
  - 対象は概念
  - 上位下位関係による体系化、様々な概念間の関係、公理による定義
- シソーラス
  - 対象は語
  - 上位下位関係といくつかの関係
- タキソノミー
  - 対象は語
  - 上位下位関係による体系化
- 概念辞書
  - 対象は語と概念
  - 概念間における上位下位関係といくつかの関係、概念-語の関係
- ボキャブラリ

# WordNet

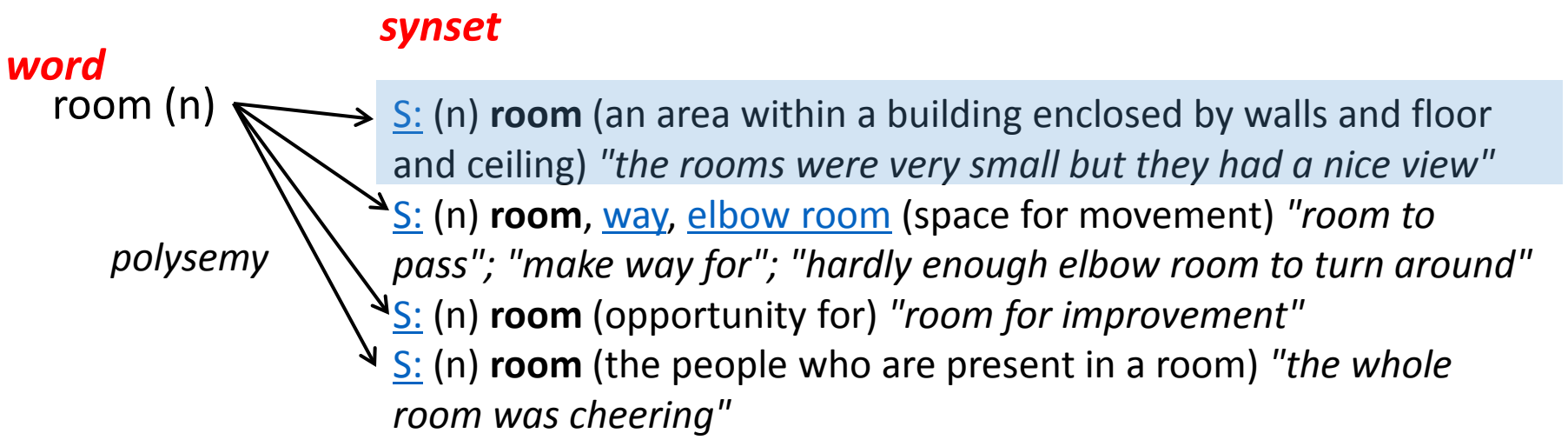
- A lexical reference system
  - "Link-based electronic dictionary" http://www.cogsci.princeton.edu/cgi-bin/webwn

| Pos | Unique Strings | Synsets | Word-Sense Pairs |
|---|---|---|---|
| Noun | 117,798 | 83,115 | 146,312 |
| Verb | 11,529 | 12,767 | 25,047 |
| Adjective | 21,479 | 18,156 | 30,002 |
| Adverb | 4,481 | 3,621 | 5,580 |
| Total | 155,287 | 117,659 | 206,941 |

- Synset Relations
  - synonym
  - hypernym/hyponym (is-a)
  - holonym/meronym (part-of)


S: (n) entity (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))
S: (n) physical entity (an entity that has physical existence)
S: (n) object, physical object (a tangible and visible entity; an entity that can cast a shadow)
S: (n) whole, unit (an assemblage of parts that is regarded as a single entity)
S: (n) artifact, artefact (a man-made object taken as a whole)
hypersym(is-a)
S: (n) structure, construction (a thing constructed; a complex entity constructed of many parts)
hypersym(is-a)
S: (n) area (a part of a structure having some specific characteristic or function)
hypersym(is-a)
S: (n) room (an area within a building enclosed by walls and floor and ceiling)
hyposym(is-a)
hyposym(is-a)
S: (n) anechoic chamber (a chamber having very little reverberation)
S: (n) anteroom, antechamber, entrance hall, hall, foyer, lobby, vestibule (a large entrance or reception room or area)
S: (n) back room (a room located in the rear of an establishment; usually accessible only to privileged groups)
S: (n) …

# WordNet Synset



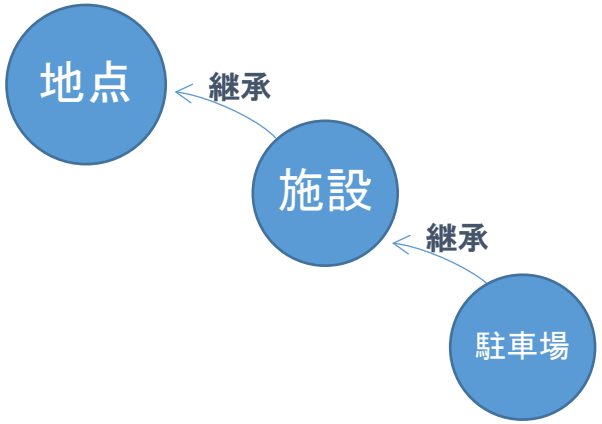
# EDR日本電子化辞書

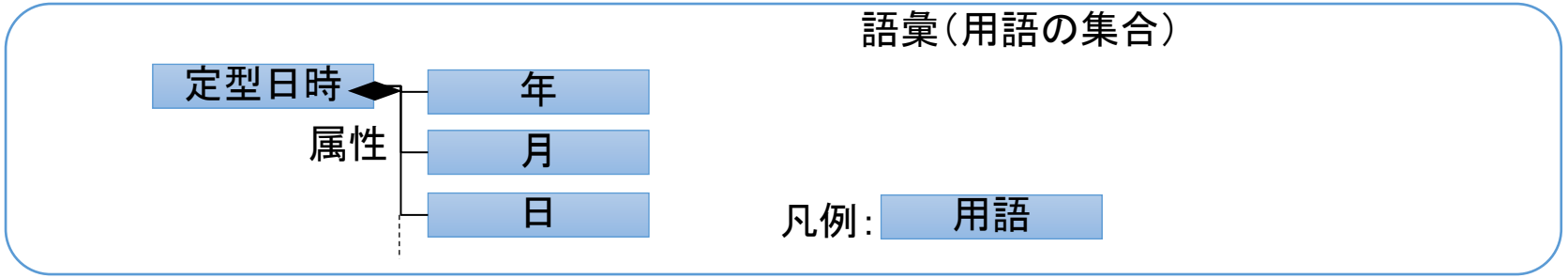
- 単語とその意味を示す概念の体系をつくる
  - 単語辞書
    - 日本語単語辞書 ・・・・・・・・・・・ 27万語
    - 英語単語辞書 ・・・・・・・・・・・・ 19万語
  - 対訳辞書
    - 日英対訳辞書 ・・・・・・・・・・・・ 23万語
    - 英日対訳辞書 ・・・・・・・・・・・・ 16万語
    - 日中対訳辞書 ・・・・・・・・・・・・ 23万語
  - 概念辞書
    - 概念体系辞書･概念記述辞書 ・・・・・ 41万概念
  - 共起辞書
    - 日本語共起辞書 ・・・・・・・・・・・ 90万句
    - 英語共起辞書 ・・・・・・・・・・・・ 46万句
  - 専門用語辞書（情報処理）
    - 日本語専門用語単語辞書（情報処理）・・ 11万語
    - 英語専門用語単語辞書（情報処理）・・・ 7万語
    - その他（概念体系、対訳、共起の各辞書を含む）
  - EDRコーパス
    - EDRコーパス ・・・・・・・・・・・ 20万文
    - 英語コーパス ・・・・・・・・・・・・ 12万文

```
=== {鍋[ナベ]}鍋という器　101bdf ===
+- 概念 3aa966
 +- ものごと　3d017c
  +- もの　　444d86
   +- 具体物 30f6ae
    +- 静物 4444c4
     +- 機能で捉えた具体物 3aa92f
        &- 器具　30f6f0
        &
        &- 器具　30f6f0
         +- 入れ物　　30f6f8
          +- 機能で捉えた入れ物 4446df
           +- 火にかけて加熱料理するための容器 4446ec
            *- {鍋[ナベ]}鍋という器　　101bdf
```

# IMI共通語彙の特徴

- 構造をもった概念辞書
  - 用語
    - 用語：個別の概念
    - 用語の表記：用語が使われるときの文字列
  - 用語の構造
    - 上位下位関係（継承関係）
      - 下位の用語は上位の用語を特殊化したもの
    - 属性関係
      - 一つの用語はいくつかの用語とそれによって指し示されるもので記述される



# IMI共通語彙の特徴

- 構造の転写
  - 属性関係をもった用語をどう表現するか？⇒シリアライズ
  - 構造をもった表現（XML, RDF)：構造をそのまま転写
  - 構造をもたない表現（自然言語）：一定の規則で変換
    - 規則を明示化して可逆性を確保

| 項目 | データタイプ | 値 |
| --- | --- | --- |
| ex:観光情報 | ex:観光情報型 | |
| ic:施設 | gf:宿泊施設_施設型<br>(extends gf:施設型<br>extends ic:施設型) | |
| @s:id | | FAC001 |
| ic:地点_名称 | ic:名称型 | |
| ic:名称_表記(日本語) | ic:テキスト型 | ○○旅館 |
| ic:名称_表記(カナ) | ic:カタカナテキスト型 | ○○リョカン |
| ic:名称_表記(英語) | ic:テキスト型 | OO hotel |
| ic:地点_場所 | ic:場所型 | |
| ic:場所_住所 | ic:住所型 | |
| ic:住所_表記(定型) | ic:定型住所型 | |
| ic:定型住所_国 | ic:テキスト型 | 日本 |
| ic:定型住所_都道府県 | ic:テキスト型 | 群馬県 |
| ic:定型住所_市区町村 | ic:テキスト型 | ○○市 |
| ic:定型住所_町名 | ic:テキスト型 | ○○町 |
| ic:定型住所_丁目 | ic:テキスト型 | 111 |
| ic:定型住所_番地 | ic:テキスト型 | 1 |
| ic:定型住所_号 | ic:テキスト型 | |
| ic:住所_郵便番号 | ic:テキスト型 | 377-xxxx |

## 型用語の属性用語の…の属性用語

**施設の名称の表記(日本語),**
**施設の名称の表記(カナ),**
**施設の名称の表記(英語),**

**…,**
**施設の場所の住所の表記(定型)の国,**
**施設の場所の住所の表記(定型)の都道府県,**

**…,**
**施設の場所の地理座標の座標参照系,**
**施設の場所の地理座標の緯度,**
**施設の場所の地理座標の経度,**

**…,**
**施設の建物,**
**施設の関連施設の施設(1),**
**施設の関連施設の施設(2),**
**施設の関連施設の役割,**

# IMI共通語彙の特徴

- 典型（プロトタイプ）としての概念
  - 共通語彙と利用者語彙
    - 共通語彙は共通かつ頻繁に概念（用語）について構造を典型例として示すもの
    - 典型（プロトタイプ）
      - 用語の構造を省略したり、追加することもできる。
      - ただし、変換規則を明示して可逆性を確保

# IMI共通語彙基盤の環境

- 語彙の階層
  - コア語彙、ドメイン共通語彙、ドメイン語彙、利用者語彙

- 語彙の利用
  - データ交換
  - データ変換

- 語彙のメンテナンス
  - 語彙の改訂プロセス

# まとめ

- IMI共通語彙とは何か
  - 概念辞書
    - 構造をもった概念辞書
    - 構造の転写による表記
    - プロトタイプとしての概念
  - 共通語彙基盤
    - 語彙の階層
    - 語彙の利用
    - 語彙のメンテナンス