森羅

SHINRA

理化学研究所　革新知能統合研究センター

「拡張固有表現＋Wikipedia」構造化データ

「構造化知識を使った言語処理応用」ワークショップ
2023年1月18日

http://shinra-project.info/

「森羅」ホームページ

構造化知識を使った言語処理応用」ワークショップ

本日のスケジュール

14:00 オープニング
14:05 森羅2022最終報告会
　　14:05 タスク説明、結果報告
　　14:25 参加システムの紹介
　　（15:00 休憩）

15:10 パネルディスカッション
　「構造化知識を使った言語処理応用」
　　・井之上直也先生（北陸先端科学技術大学院大学）
　　・河原大輔先生（早稲田大学）
　　・山田育矢様（株式会社Studio Ousia）
　　・中山功太（理研AIP/筑波大学）
　　・関根聡＜司会＞（理研AIP)

16:10 森羅2023の紹介
16:25 クロージング
　　（16:30 閉会）

森羅プロジェクトの最終目標

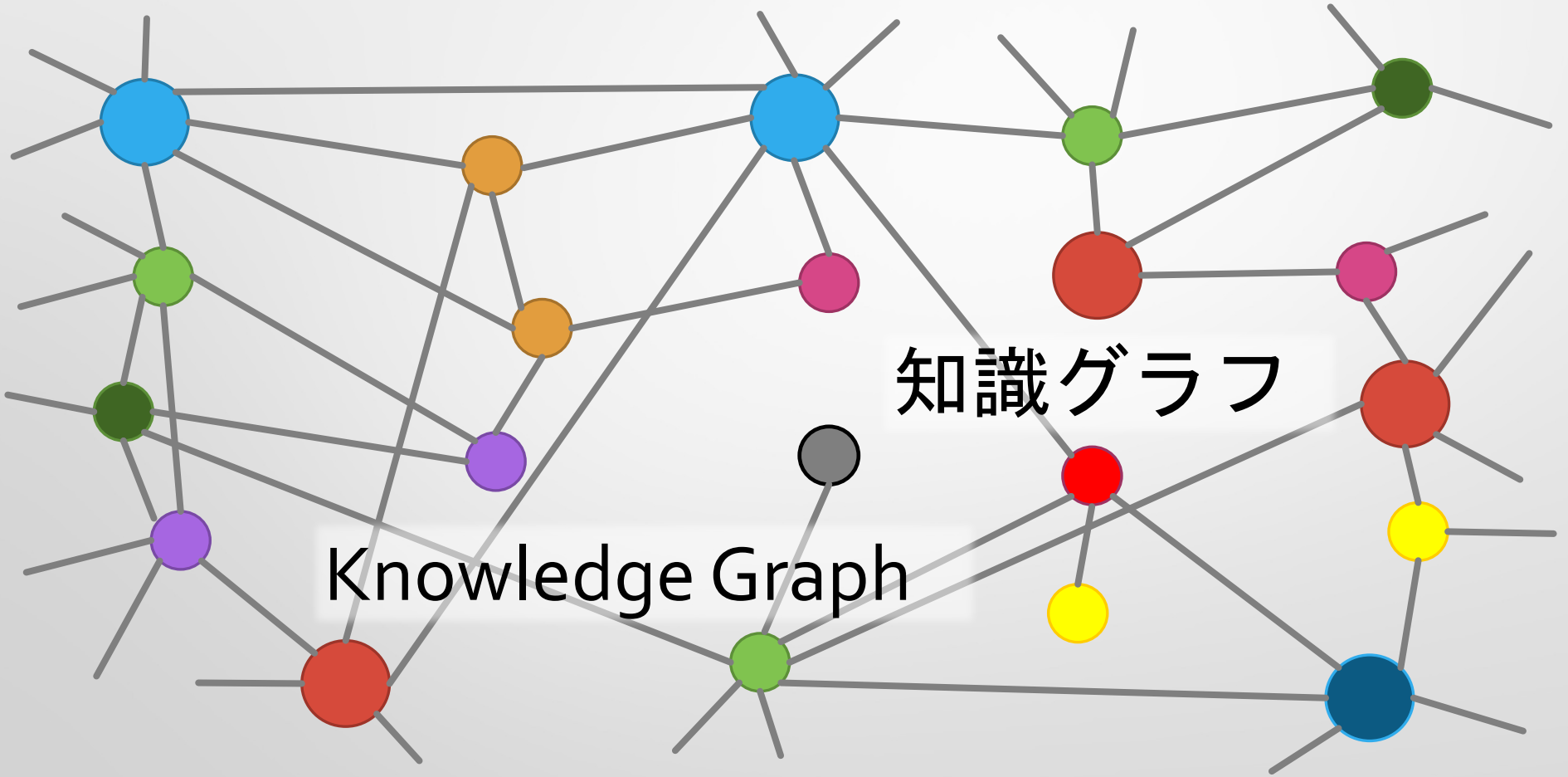
信頼される人工知能

単に答えを示すだけではなく、
答えの根拠を人が理解できる
説明の形で提示する。

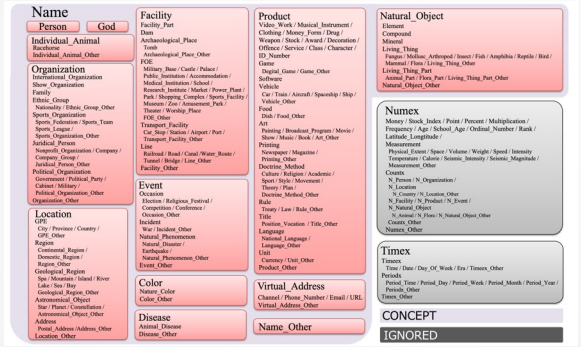
人工知能の普及に新しい展開

構造化された世界知識
「森羅」

森羅 ＝ Knowledge Graph 知識グラフ



知識グラフを作ろう！



Wikipediaの記事を拡張固有表現に分類し
定義された属性値を抽出して構造化、リンク作成

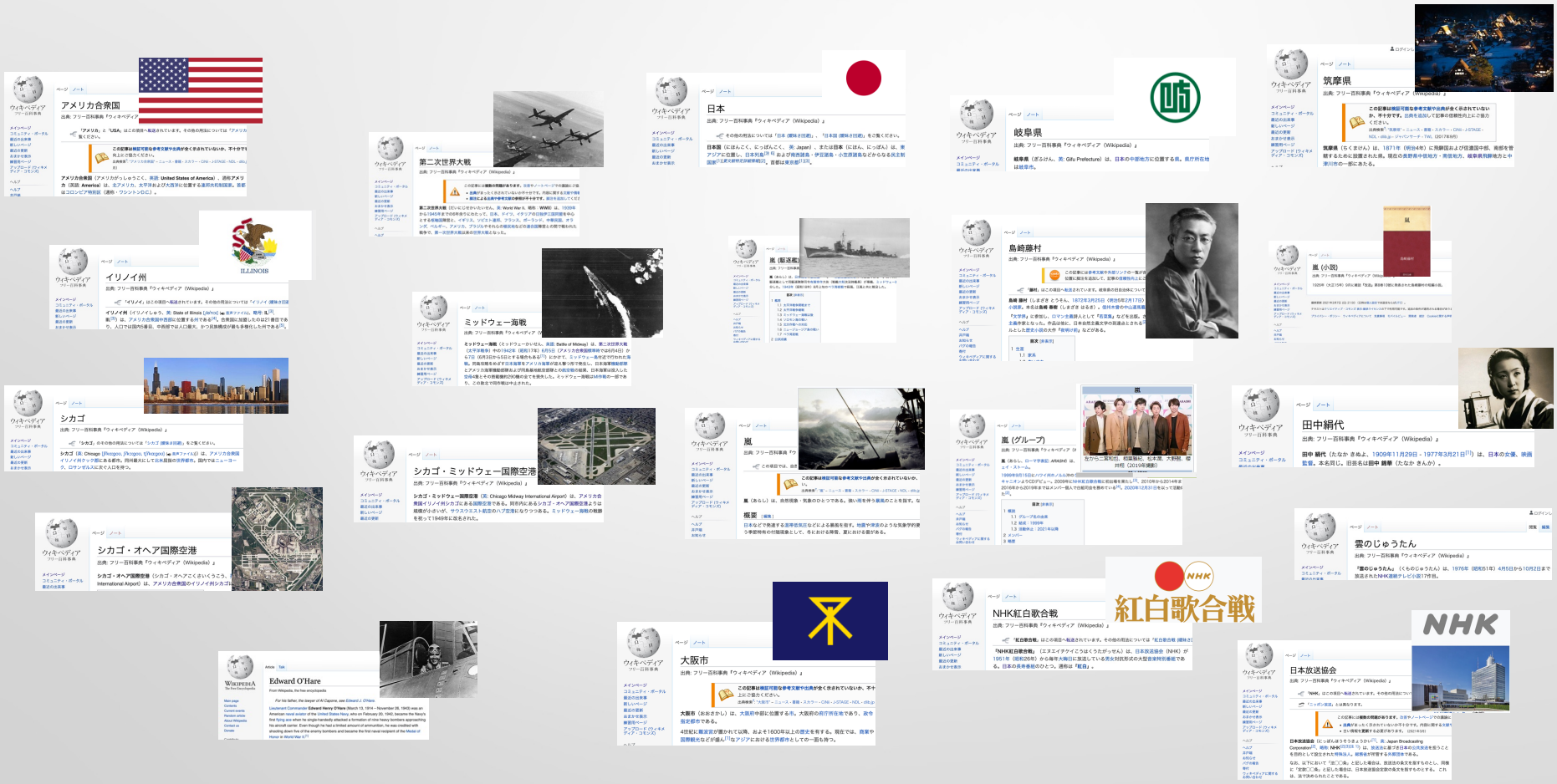
Resource by Collaborative Contribution

Since 2017

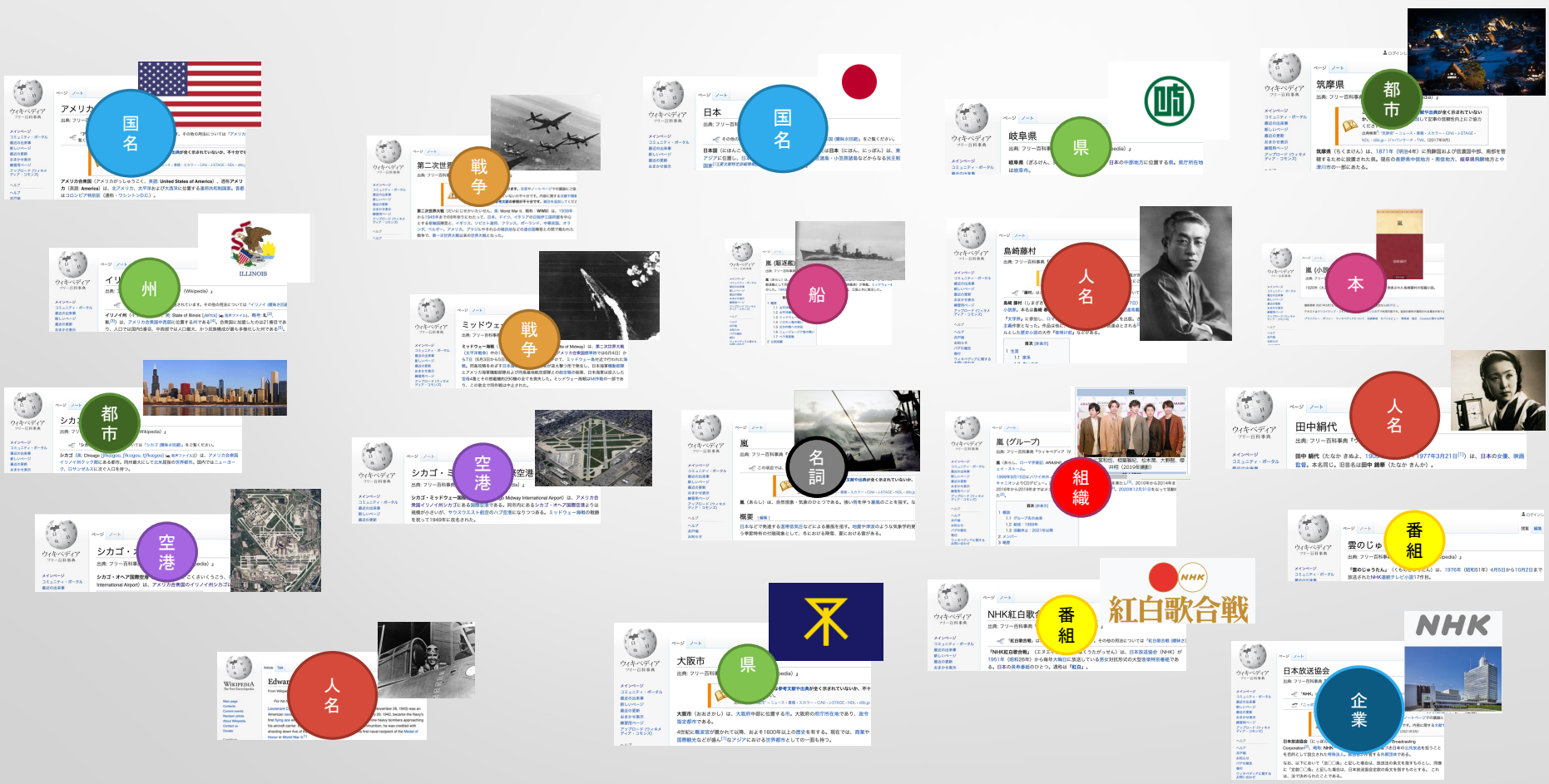
評価WSを通して協働で知識作成

Wikipediaページ

（日本語は100万、英語は600万）

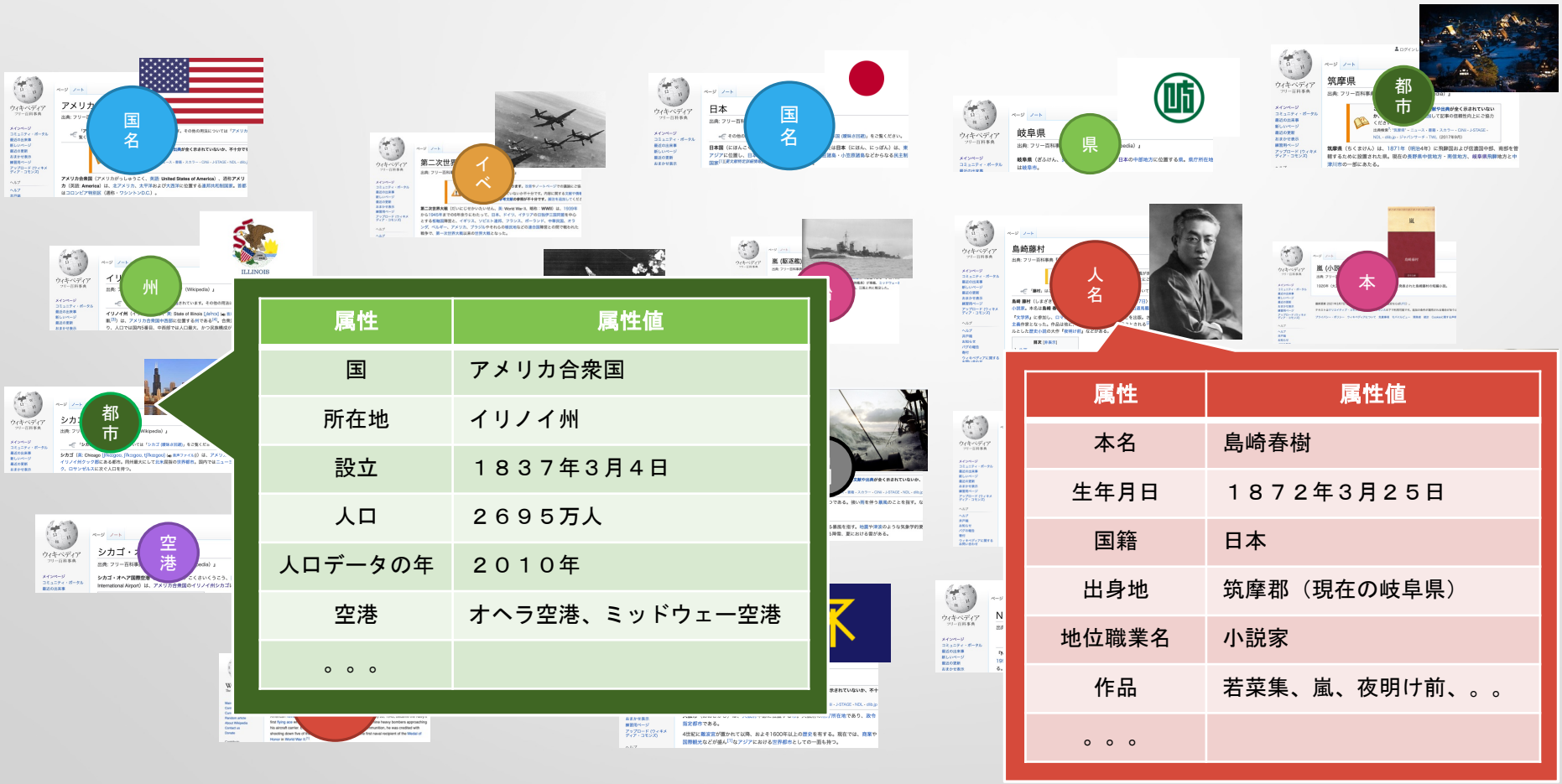


ステップ１（分類）



森羅：Wikipedia構造化プロジェクト

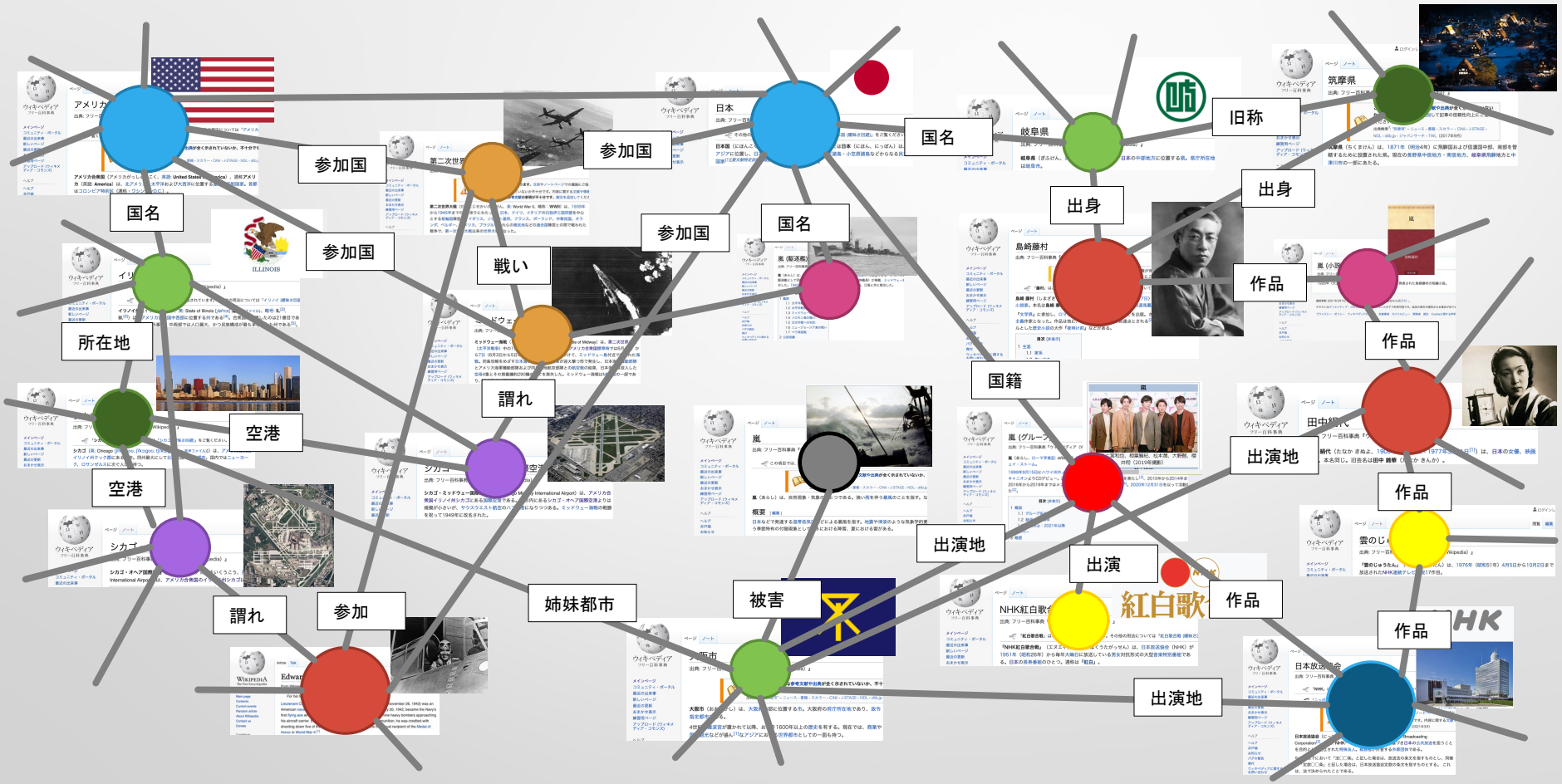
ステップ２（属性値抽出）



属性	属性値
国	アメリカ合衆国
所在地	イリノイ州
設立	１８３７年３月４日
人口	２６９５万人
人口データの年	２０１０年
空港	オヘラ空港、ミッドウェー空港
ｏ ｏ ｏ	

属性	属性値
本名	島崎春樹
生年月日	１８７２年３月２５日
国籍	日本
出身地	筑摩郡（現在の岐阜県）
地位職業名	小説家
作品	若菜集、嵐、夜明け前、。。
ｏ ｏ ｏ	

Knowledge Graph
知識グラフ



森羅：Wikipedia構造化プロジェクト

協働による知識構築

Resource by Collaborative Contribution (RbCC)

- 評価型ワークショップを実施
- 単に性能を競い合うだけではない
- **参加システムがリソース作成に直接貢献**
  - 例えば、１０チーム中８チームが正しいといったものは正しいとする　(Ensemble Learning)
  - 適切な人手チェックを入れデータを拡張 (Active Learning)
  - 拡張した教師データーで再度タスクを実行 (Bootstrapping)
- すべての出力データは参加者で共有する
- 統合されたデータは一般に公開する

知識グラフがあると

物知り博士

２つの空港がある米国の都市で、両方とも第二次世界大戦に由来した名前がついているのはどこでしょう？

オヘア空港とミッドウェー空港を持つシカゴ！

雑談対話

おじいちゃんの好きな田中絹代が出演した1957年の「嵐」って映画の原作は、島崎藤村の小説なんだって

情報アクセス　教育支援　営業支援　認知症予防　介護福祉

他にも。。。

多言語情報アクセス

特定応用展開

観光　外国語教育　ビジネス

特許検索・分析　法律文書解析

オンライン医療　健康自己管理

森羅：Wikipedia構造化プロジェクト

http://shinra-project.info/

「森羅」ホームページ

過去に行った９タスク

タスク名	タスク	言語	詳細
2018	属性値抽出	日本語	5カテゴリー
2019	属性値抽出	日本語	35カテゴリー
2020-JP	属性値抽出	日本語	78カテゴリー
2020-ML	分類	30言語	
2021-LinkJP	リンク	日本語	7カテゴリー
2021-ML	分類	30言語	
2022	分類 属性値抽出 リンク	日本語	全(178)カテゴリー

森羅2022の現状

- 参加者

タスク	リーダーボード	全件予測
分類	５９	５
属性値抽出	３	—
リンク	２	—

- 属性値抽出、リンクタスク
  - 分類同様のワークショップ？
  - 参加の敷居を下げる必要？

改善アイデア

1. タスク形式について
   - 属性値抽出、リンクタスクでは、全件のWikipedia記事を対象とするのではなく、特定のカテゴリーなど限定された記事を対象にすることを可能とする
   - 前段階の間違えを含んだデータを対象にEnd-to-Endタスクだけではなく、前段階の正解データに基づいて参加できる仕組みを作る
2. 提供するデータについて
   - 機械学習がより精度高くなるように学習データを拡充する
   - 多くの開発データを配布し、参加者が詳細な分析をできるようにする
   - データの一貫性、前処理の簡易化を行う
3. デモタスク
   - 森羅データを利用した応用システムのデモタスクを用意する

準備中: 森羅 Public API

- 森羅にて作成された全データの利用が可能（RbCC）
  - 分類（多言語）
  - 拡張固有表現　分類・属性定義
  - Wikipediaページの拡張固有表現抽出結果
  - 拡張固有表現のリンク結果

```
$ curl -H "Authorization: Bearer $TOKEN" \
  https://api.shinra-project.info/categories/7bf8f066-779b-415c-8d2e-86a641850f4f | jq '.'
{
  "id": "7bf8f066-779b-415c-8d2e-86a641850f4f",
  "ene_id": "1.8.3.3",
  "name": {"en": "Competition", "ja": "競技会名"},
  "definition": {
    "en": "A name of a competition or game of a sport, or an event competing the superiority of any skill.",
    "ja": "スポーツなどの競技大会の名前。コンクールなどのスポーツ以外の競技会を含む。"
  },
   . . .
  "attributes": [
    {"id": "8b330619-88d6-41c4-a999-8fe82509f217", "name": "読み"},
   . . .
  ],
  "ene_version": "9.0"
}
```

森羅データセット

タスク	学習データ	システム出力
日本語分類	「2022分類教師データ」 ページ数：920,444	*森羅2022結果を準備中* ページ数：1,286,205 システム数：5(?)
多言語(30言語) 分類	「2020多言語分類教師データ」 言語数:30 ページ数：5,029,617	「2020/2021多言語出力データ」 言語数:30 システム数:12 ページ数：32,555,929
属性値抽出	「2022属性値抽出教師データ」 ページ数：19,711 属性種類数：1,671 属性値延数：910,567	「2020属性値出力データ」 システム数：13 カテゴリー数：35 属性値数：6,089,547
リンク	「2021リンク教師データ」 7カテゴリー、350ページ、7,284リンク 「2021リンク開発データ」 7カテゴリー、706ページ、13,887リンク 「2022リンク教師データ」 178カテゴリー、1397ページ、59,429リンク	*森羅2022結果を準備中*

森羅2023への参加を是非

- コミットメントのために
  - 理研での学生バイトなどの用意
  - 研究成果の学会発表、卒論、修論
- 興味ある方は、
  - 森羅slackへの参加を（こちら）！
  - 関根(satoshi.sekine@riken.jp)まで

森羅HP
http://shinra-project.info

Slack参加ページ
http://shinra-project.info/shinra2022/shinra2022_slack_invite