iRODS を用いたデータ管理システムの導入

○飯田 好美

高エネルギー加速器研究機構 共通基盤研究施設 計算科学センター

概要

KEK 共通計算機システム(KEKCC)では大容量ストレージシステムとして HPSS を導入し、J-PARC 実験のデータの保管場所として提供している。また、2009 年 3 月の KEKCC 更新に伴いデータ管理システムの一つとして iRODS を導入し、KEK と J-PARC 間の高速なデータ転送、サイト間にまたがるストレージの仮想化、異なるストレージへのインターフェイスの統一を実現するシステムとして提案している。

ここでは KEKCC の iRODS システムについて報告する。

1 はじめに

高エネルギー加速器研究機構（以下、「KEK」という）では、大型加速器を中心とする研究施設を利用して、高エネルギー物理学、物質科学、生命科学などの幅広い研究が行われている。計算科学センターでは、これらの研究活動に必要な計算機・ネットワーク環境の導入、整備、運用を行っている。

KEK 共通計算機システム（以下、「KEKCC」という）は主に大規模な計算機サーバ群とストレージシステムから構成されており、素粒子原子核実験、放射光実験、中性子実験、加速器開発、理論計算等の様々な研究ニーズに応じたシステムを提供するものである。特に、KEK と日本原子力研究開発機構（JAEA）の共同プロジェクトである J-PARC 実験ではペタバイトオーダーでのデータ量が見積もられており、KEKCC の主要ユーザとして位置づけられている。しかし、J-PARC の実験施設は茨城県東海村にあり、KEK から 60km ほど北部に位置している。実験によって検出されたデータは、一旦各実験グループが保有するストレージシステムに蓄積され、その後 KEKCC へと転送する必要がある。

KEKCC では KEK と J-PARC 間の高速なデータ転送、サイト間での異なるストレージの管理などを解消するために iRODS (the Integrated Rule-Oriented Data System)[1]による論理的分散システムの構築を行い、その成果を報告する。

2 KEK 共通計算機システム（KEKCC）

KEKCC は 2009 年 3 月にシステム更新を行い、それまで KEK イントラネットワークに設置していたシステムを KEK-FW の外に出し、KEKCC 独自の FW 内のネットワークに接続した。国内外の研究機関と連携して研究を進めている LCG、NAREGI、iRODS を含む GRID システムは、その性質から非常に多くのサイトに対して多種のポートを開放する必要があるが、独自の FW を構築することにより KEK イントラネットワークのセキュリティレベルを落とすことなく、KEK 外のネットワークとの相互接続に対して柔軟な対応が可能となった。

図 1 は KEKCC の構成図である。ワークサーバはエンドユーザが直接ログインして利用するサーバであり、ジョブを作成したり、計算サーバにバッチジョブを投入したりする。磁気ディスクシステムはワークサーバ、計算サーバのホーム領域の提供や、高速なストレージ領域の提供を行い、大容量ストレージシステムとしては HPSS を備えている。

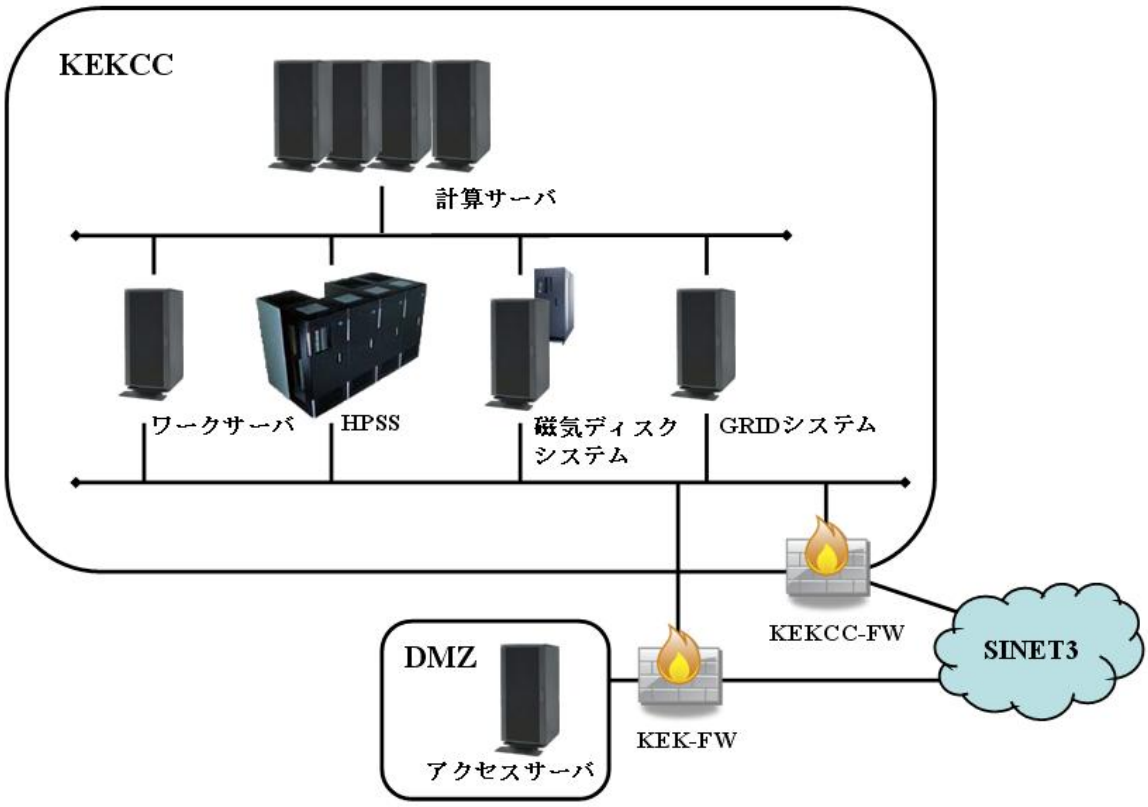

図 1.　KEKCC 構成図

2.1　大容量ストレージシステム

KEKCC では大容量ストレージシステムとして HPSS（High Performance Storage System）を導入しており、これは 10TB の磁気ディスクと最大容量 3PB のテープ装置で構成された階層型ストレージシステムである。

HPSS へは図 2 にように主に 6 種類のインターフェイスを提供している。

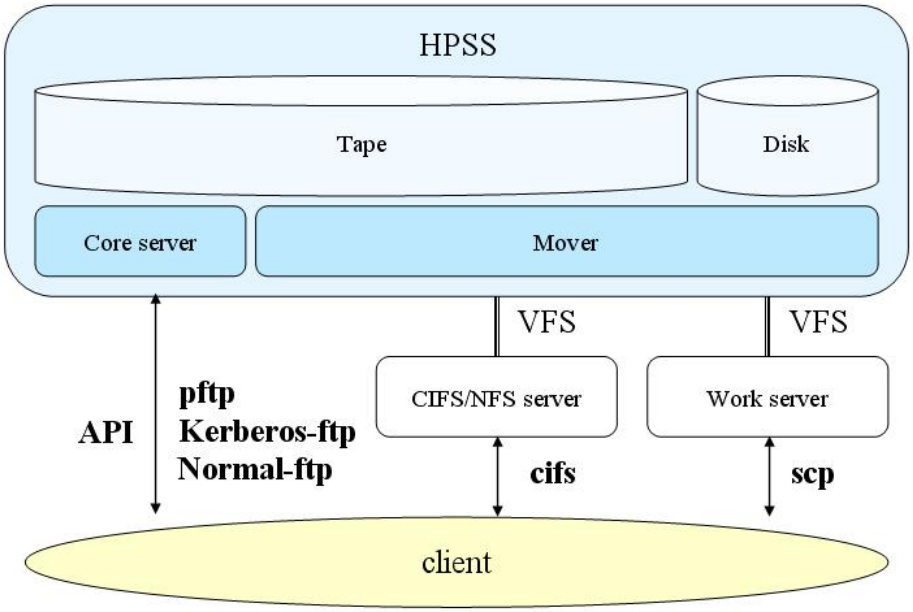

図 2.　HPSS インターフェイス

API は HPSS パッケージを導入することで使用できるようになるツールで、C/C++プログラムからデータ転送を行う ClientAPI とインタラクティブにデータ転送、ファイル操作を行える HPSS コマンドがある。クライアントは HPSS のコアサーバへ接続され、ムーバーから直接データ転送を行う。転送速度は速いが、1 セッションあたりの HPSS へのコネクション数が多いためクライアントが増えると接続できなくなる可能性があること、Linux 系しかサポートしてないことなどが問題点として挙げられる。

pftp も API 同様に HPSS パッケージを導入することで使用できるようになるツールで、インタラクティブにデータ転送を行う。クライアントは HPSS のコアサーバへ接続され、データサイズによってコアサーバ経由またはムーバーから直接データ転送が行われる。転送速度は速いが API と同様の問題点がある。

Kerberos-ftp、Normal-ftp は Linux でサポートされている ftp コマンドで、コアサーバ経由でインタラクティブにデータ転送を行う。Normal-ftp は認証パスワードがプレーンテキストでネットワークを流れるため、他に利用できるツールがない場合を除いては通常 Kerberos-ftp を推奨している。Kerberos-ftp は HPSS への認証方式として Kerberos を使用した ftp で、クライアントマシンに Kerberos を導入する必要がある。これらのクライアントは KEKCC や KEK 内からの転送速度は速いが、転送距離が伸びると速度は落ちる。

cifs は Windows のファイル共用サービスとして利用されている SMB を拡張したもので、Windows だけでなく Unix 系 OS やアプリケーションからも利用可能なサービスである。CIFS/NFS サーバは HPSS を VFS でマウントしており、クライアントは CIFS/NFS サーバに接続することで HPSS をファイルシステムとして操作することができる。CIFS のクライアントとしては GUI アプリケーションなども多く、直感的に使うことが可能だが、転送速度は遅い。

scp は Unix 系の OS にはデフォルトで導入されており、インタラクティブにデータ転送を行うことができる。ワークサーバには HPSS が VFS でマウントされているため、クライアントはワークサーバに scp しているのと同じように HPSS を利用することができる。scp はセキュアでよく知られた転送ツールだが転送速度は遅い。

これらのインターフェイスは KEKCC の計算サーバや KEK イントラネットワークから HPSS のデータを利用する際には良いが、転送距離が長く、セキュリティポリシーの異なる J-PARC とのデータ通信には向かない点が多い。そこで KEKCC では長距離転送にも強く、異なるストレージシステムを一元的に管理することができる iRODS の導入を行った。

3 iRODS とは

iRODS は DICE(the Data Intensive Cyber Environments)グループが中心となって開発されたルール指向のデータグリッドソフトウェアシステムで、ネットワークで接続された複数のストレージリソースを 1 つのファイルシステムとして提供することが可能である。iRODS では ICAT(iRODS metadata catalog)がアカウント、ストレージリソース、データファイルなどの管理を行っており、1 つの ICAT と 1 つまたは複数の iRODS サーバによってシステムが構成される。また、iRODS の入出力では並列転送をサポートしており、ファイルサイズに応じてスレッド数が自動的に設定される。これにより高速なデータ転送が期待できる。なお、iRODS システム内では、ファイルのことを"データオブジェクト"、ディレクトリのことを"コレクション"と呼ぶ。

3.1 iRODS コンポーネント

図 3 は iRODS コンポーネントの概要を表した図である。

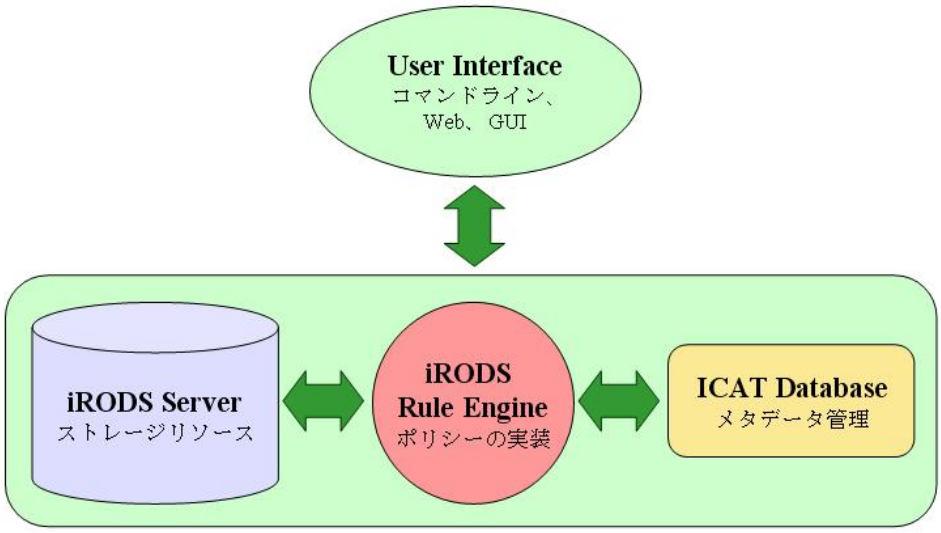

図 3.　iRODS コンポーネント概略図

ユーザーインターフェイスとしては、コマンドライン、Web アプリケーション、GUI アプリケーションなどがあり、iRODS のユーザ認証やファイル操作、メタデータの作成など、クライアント機能を提供する。基本的には iRODS パッケージやクライアントツールのインストールが必要だが、Web アプリケーションは Web ブラウザさえあれば特別な設定が不要である。

iRODS サーバはクライアントからのリクエストを受け付け、ICAT で照合を行い、ルールに基づいてリクエストを実行する。1 つの iRODS システム内に複数の iRODS サーバが存在する場合でも ICAT に問い合わせを行う iRODS サーバは 1 つだけで、それ以外のサーバは必ずその iRODS サーバを経由して照会を行うことになる。また、iRODS で利用するストレージリソースは必ず iRODS サーバに接続される必要がある。

iRODS ルールエンジンは全ての iRODS サーバに実装されており、システムのポリシーを設定しているものである。ルールエンジンについては次項で詳細を説明する。

ICAT は iRODS 内で使用するユーザ、ストレージ、データ、メタデータ等の管理を行うデータベースである。iRODS システム内で使用する論理名と、iRODS サーバ上に存在する実体とのマッピングを行う。iRODS サーバ上では実ユーザは iRODS の起動ユーザのみであり、実ファイルの作成などは全てこのユーザが作成するので、iRODS 内で設定したアクセスコントロールは iRODS 内でのみ通用するものである。

3.2　ルールエンジン

iRODS ルールシステムの核となっているのがルールエンジンである。これは全ての iRODS サーバに実装されており、システムにポリシーを設定することができる。ルールエンジンは設定した条件に基づき、事前に定義されたマイクロサービスを呼び出すことができる。

マイクロサービスとはあるタスクを実行するための小さなファンクションである。データオブジェクトのコピー、削除、ダウンロードや、コレクションの作成、削除など、デフォルトで用意されているマイクロサービスも多い。また、ユーザが独自のマイクロサービスを作成することも可能である。

ルールは状態と動作を設定することができ、状態とはルールを実行するための条件、動作とは呼び出すマイクロサービスの種類と順番である。条件の設定としては時間、実行頻度、データオブジェクトの名前、ユーザ、ストレージなどを使用することができる。また、動作で定義したマイクロサービスは書かれた順に実行されるため、複数のマイクロサービスを指定することで複雑な動作をおこなうことも可能である。

4　KEKCC の iRODS システム

図 4 は KEKCC 内に構築した iRODS システムの構成図である。

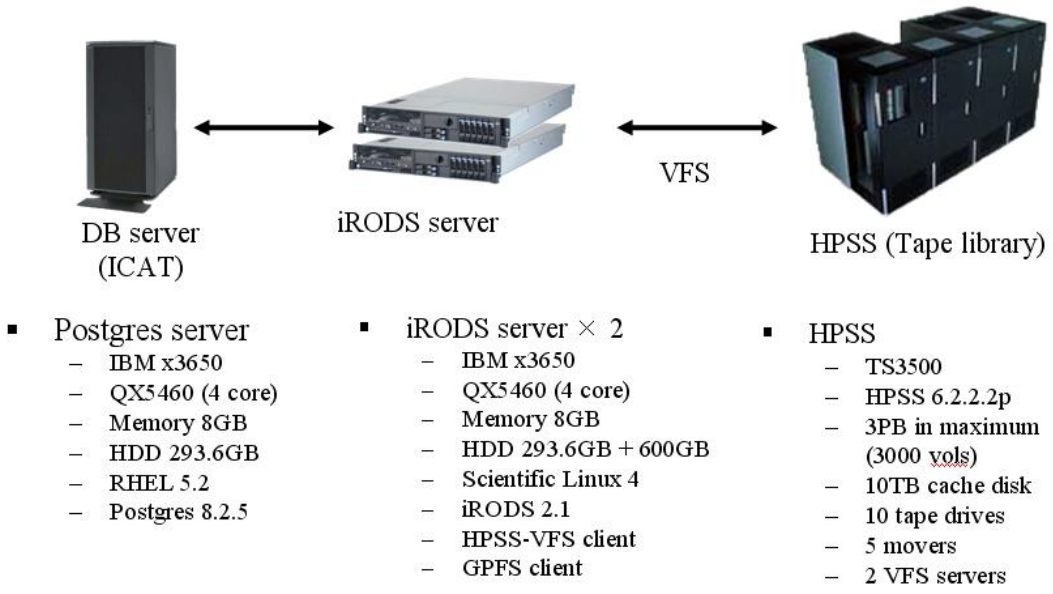

図 4.　KEKCC iRODS 構成図

KEKCC では iRODS サーバとして IBM x3650 を 2 台設置し、1 台は稼動機、もう 1 台は予備機とした HA 構成を組んでいる。iRODS システムではサーバの持つ IP アドレス以外にサービス IP アドレスが存在し、iRODS サーバへのアクセスはサービス IP を指定することになる。稼動機の障害発生時にはサービス IP を手動で予備機へスイッチすることでサービスの停止を最小限に抑える。

ICAT は iRODS サーバとは物理筐体を分けた DB サーバにインストールされている。iRODS のデフォルトの設定では、ICAT は iRODS サーバにインストールされるようになっているが、KEKCC では iRODS における ICAT の重要性を考え、DB 専用のサーバに搭載することとした。

ストレージリソースとしては 600GB の外付け HDD と HPSS が用意されている。HPSS は iRODS サーバに VFS(Virtual File System)マウントされており、HPSS の領域をあたかもローカルファイルシステムとして扱うことができる。

4.1　転送性能

iRODS は並列転送をサポートしており、以下の計算式で得たスレッド数を自動的に設定しデータ転送を行う。

- スレッド数 = ファイルサイズ(MB) ÷ 1 スレッド当たりのサイズ(MB) + 1

1 スレッドあたりのサイズはサーバ毎に設定することが可能で、デフォルトは 32MB となっている。KEKCC ではデフォルト値を採用しているため、32MB より大きなサイズのファイルを転送する際に並列転送が使われることになる。また、1 つのファイルに対して同時に使用することができる最大スレッド数も設定することができ、iRODS では最大 16 本の並列転送までサポートしている。デフォルトは 4 本となっているが、KEKCC では最大の 16 本まで使用するように設定している。

これらの設定を行ったうえで、J-PARC(東海村)に設置したクライアントから KEK(つくば)に設置した HPSS までの転送速度を測定した。なお、クライアントマシンのスペックは表 1 の通りである。転送に使用したデータは 1GB のファイルで、コマンドラインインターフェイスである iput と iget の 2 種類のコマンドを使用し

た。iput はローカルファイルを iRODS 内にアップロードするコマンドで、クライアントマシンから iRODS リソースである HPSS へファイルをアップロードすることで、HSPP への書き込み速度を測定した。iget は iRODS 内にあるデータオブジェクトをローカルファイルにダウンロードするコマンドで、HPSS に保管されているデータオブジェクトをクライアントマシンへダウンロードすることで、HPSS の読み出し速度を測定した。

表 1. クライアントマシンスペック

CPU	Intel Xeon 3.0GHz 8 コア
LAN	Gigabit Ethernet
Memory	4 GB
OS	CentOS release 5.2

表 2 はその測定結果である。比較として、八代氏が同クライアントマシンを使用し他のインターフェイスで転送性能を測定した結果[2]を載せる。

表 2. KEK-JPARC 間の HPSS 転送性能

ツール	HPSS 書き込み（J-PARC→KEK）	HPSS 読み出し（KEK→J-PARC）
iRODS	43MB/s	40MB/s
pftp	26MB/s	17MB/s
scp	24MB/s	4MB/s

この結果から、iRODS が HPSS のインターフェイスとして有効であることがわかる。

4.2 ユーザーインターフェイス

iRODS のユーザーインターフェイスとしては複数のツールが用意されているが、現在 KEK では以下の 3 種類をサポートしている。

- i-Commands (iRODS コマンドラインインターフェイス)
- JUX[3] (Java GUI)
- Davis[4] (Web アプリケーション)

i-Commands は前項の転送速度測定でも使用したツールで、Linux と Windows で使用することができるコマンドラインインターフェイスである。Unix 系のコマンドの先頭に”i”を付けた形のコマンドが多い。iRODS の開発グループがメインでサポートしており、最も軽く、転送速度も早いツールである。iRODS のパッケージからインストールすることができ、サーバからクライアントマシンへの接続は発生しないので、クライアントは OUT 方向のポートが開いていれば良い。

JUX(Java Universal eXplorer)は Windows、Mac、Linux にインストールすることができる Java アプリケーションである。iRODS の共同開発機関の 1 つである CC-IN2P3 で開発され、iRODS ファイルシステムを視覚的に操作することができる。Windows エクスプローラー同様、ウィンドウの左側にはフォルダ(コレクション)ツリーが、右側には選択したフォルダ(コレクション)の中が一覧表示される。1 つのウィンドウ内にローカルファイルシステムと iRODS ファイルシステムが混在するため、ドラッグ&ドロップでローカルディスクと iRODS 間のファイル転送が可能である。クライアントマシンに Java をインストール後、JUX のパッケージをインストールすることで使用可能になる。

Davis(WebDAV-iRODS/SRB gateway)はクライアント側で設定の必要がない Web アプリケーションである。IE、Firefox、Safari などの Web ブラウザから Davis のサーバにアクセスし、アカウント名、パスワードによって認証を行うことで iRODS ファイルシステムに接続できる。使用するポートも https のみであるため、外部との接続が制限された研究機関、大学等からも使用可能であると期待できる。

4.3 ルール

KEKCC では J-PARC 実験のデータを HPSS へ転送するためのポリシーを作成し、そのルールのテストを行っている。

J-PARC 実験で採取されたデータはすぐに J-PARC 施設内で解析されるため、採取直後は J-PARC 施設内からのアクセスが圧倒的に多い。そのため、この期間は J-PARC のストレージに保管する必要がある。しかし一定期間経過後はアクセスが激減し、J-PARC 以外の研究所、大学などからのアクセスが発生する。J-PARC のネットワークは外部からのアクセスが非常に制限されているため、外部からのアクセスは難しい。また、ストレージの容量にも余裕が少ないため、長期保管用ストレージへ移行したデータはストレージから削除したい。そのため、J-PARC 実験データの移行に必要なポリシーとして、J-PARC 実験データの HPSS への移行と J-PARC のストレージのリフレッシュが挙げられる。

データの移行には iRODS のレプリカ機能が利用できる。レプリカとは iRODS ファイルシステム上は 1 つのデータオブジェクトに対して物理的に異なる複数のストレージにファイルを作成することである。複数のストレージにあるファイルを 1 つのオブジェクトとして見せるので、一方のストレージが障害などで使用できない場合もユーザには影響がない。また、ストレージのリフレッシュにはトリム機能が利用できる。トリムとは iRODS のレプリカを削除する機能で、最小レプリカ数を指定することにより誤ってデータが削除されることが防げる。これらを使い、J-PARC のストレージに保存されたデータを HPSS へレプリカし、一定期間後、J-PARC から削除するというルールを作成している。

現在は、実行頻度等の詳細設定、ネットワーク障害、ストレージ障害時の対処法、HPSS 領域の設定、実験グループ毎の設定の違いなどについて検討を行っている。

5 まとめ

現在 KEKCC では新しいデータ管理システムの 1 つとして iRODS の導入を行っている。HPSS インターフェイスとして高速なデータ転送が期待でき、J-PARC からの転送速度は約 40MB/s である。また、複数のストレージを 1 つの iRODS ファイルシステムとして提供することができるため、J-PARC と KEK のストレージを意識することなく使用することが可能である。また、ルールを作成することで、必要なデータのみを自動で HPSS へ転送することが可能である。

今後はユーザが独自のルールを作成できるよう、ユーザサポートページの充実や講習会の開催などを検討したい。

参考文献

[1] https://www.irods.org/index.php

[2] 八代茂夫、“JPARC 実験データの共通計算機システムへの転送”、平成 21 年度 KEK 技術研究会報告集、平成 22 年 3 月

[3] https://forge.in2p3.fr/wiki/jux/Jux

[4] https://projects.arcs.org.au/trac/davis/wiki/WikiStart